IMSE-cnm Instituto de Microelectrónica de Sevilla

CSIC Consejo Superior de Investigaciones Científicas — Universidad de Sevilla

Doctoral Thesis

Design of CMOS Digital Silicon Photomultipliers with ToF for Positron Emission Tomography

Author:
Franco Nahuel Bandi

Advisors:
Ricardo Carmona Galán
Ángel Rodríguez Vázquez

Sevilla 2020

IMSE-cnm Instituto de Microelectrónica de Sevilla

CSIC Consejo Superior de Investigaciones Científicas — Universidad de Sevilla

Design of CMOS Digital Silicon Photomultipliers with ToF for Positron Emission Tomography

By:
Franco Nahuel Bandi

PROPUESTA DE TESIS DOCTORAL
PARA LA OBTENCIÓN DEL GRADO DE
DOCTOR EN CIENCIAS Y TECNOLOGÍAS FÍSICAS

Advisors:
Ricardo Carmona Galán
Ángel Rodríguez Vázquez

Seville, 2020

Design of CMOS Digital Silicon Photomultipliers with ToF for Positron Emission Tomography

Doctoral Thesis
Universidad de Sevilla

Franco Nahuel Bandi
Master in Microelectronics

Advisor: Ricardo Carmona Galán
Tenured Scientist
Instituto de Microelectronica de Sevilla
Consejo Superior de Investigaciones

Advisor: Ángel Rodríguez Vázquez
Full Professor
Electronics and Electromagnetism Department
Universidad de Sevilla

Universidad de Sevilla
Consejo Superior de Investigaciones Científicas
Centro Nacional de Microelectrónica
Instituto de Microelectrónica de Sevilla

Calle Américo Vespucio, 28.
Parque Científico y Tecnológico Cartuja.
41092 Sevilla

Author e-mail: nahuel@imse-cnm.csic.es

Acknowledgements

Será imposible hacer justicia en estas líneas y agradecer a toda la gente que me apoyó y, de alguna forma u otra, hizo posible este desenlace, por lo que tal vez debería empezar pidiendo disculpas por el olvido y la brevedad. No creo que haya un solo evento, uno discreto e independiente, que haya desembocado en la realización esta tesis. Más bien, esto probablemente sea consecuencia de una integración de circunstancias, pequeñas, ingravidas y sutiles. ¡Gracias totales!

I would like to thank my advisor Dr Ricardo Carmona Galán for his mentoring and support over these years. His experience and knowledge in electronics and image sensor design was fundamental at each stage of the thesis. My gratitude must be extended to my co-advisor Dr Ángel Rodríguez Vázquez, who provided very valuable advice during the analysis and design of circuits. Their support and passion for this subtle art allowed me to grow as a designer, but more important, as a person.

También tengo que agradecer al Ministerio de Economía y Competitividad por proporcionar la financiación de mi contrato FPI (BES-2016-076255) que me permitió realizar mi tesis.

I would like to thank Dr Rafael Ballabriga Sune and Dr Michael Campbell for giving me the opportunity of doing a four-months research stay at CERN, where I learned a lot about time-to-digital converter design, fast timing applications and X-ray detectors. Also, I would like to thank Iraklis, Samuele, Cienti, Sara M. V., Jerome and Wojciech for all their invaluable help and patience during the whole stay. Specially, I am very grateful to Jose and Nuria not only for sharing their expertise in fast timing, but also for our non-technical discussions.

Quiero agradecer a Ion por sus discusiones de carácter técnico, su ayuda a la hora de dilucidar los entresijos del funcionamiento de algunos circuitos y su soporte durante la caracterización del chip. Tampoco puedo olvidar la ayuda recibida de Joaquín, Antonio R., Antonio L., Cristina y Luis durante los test en el laboratorio. No puedo dejar pasar la ocasión sin agradecer a mis compañeros

del grupo, Jorge, Juan Antonio, James, Raffaela y Marco, con los tuve la suerte de coincidir. Asimismo, me gustaría dar las gracias a todo el personal del IMSE, área técnica, informática, administración y mantenimiento, por su ayuda y por crear un espacio de trabajo acogedor. A todos ustedes ¡Gracias!

With a special mention to all the members, past and future, of the 34: José Luis, Alfredo, Andrés, Héctor, Beti, Luca, Giuseppe, Norber, Pablo, Antonio, Valentín, Juan, David, Virginia, Rubén e Isma. For those all good and bad moments that you where there, giving me advice and being patient. It was an honor to share the space with all of you during these last years.

Me gustaría aprovechar estas líneas para agradecer a todos mis profesores y maestros. En particular, a aquellos que dedicaron su tiempo y paciencia en hacerme crecer como persona y me enseñaron que la vida es el camino, no los hitos. ¡Grego, Sebastian, Jaume, Rodrigo y Eugenio muchas gracias!

Finalment, no puc deixar d'agrair als meus amics Cristian, Sergio, José, Joan, Miquel, Roberto, Naza, Inés i Juan que m'han donat el seu suport al llarg de tot aquest camí i, sempre que ho vaig necessitar, em varen donar un lloc tranquil on descansar. Gràcies a tots vosaltres per omplir la meva memòria de bons records, els meus ulls de la verdor de la muntanya i els meus pulmons de la frescor de la mar!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

a-SiPM Analog Silicon Photomultiplier

ADC Analog to Digital Converter

ADPLL All-Digital Phase Locked Loop

AQRC Active Quenching and Recharge Circuit

BGO Bismuth Germanate

CAD Computer-Aided Design

CIS CMOS Image Sensor

CsI Cesium Iodide

CT Computed Tomography

CTR Coincidence Time Resolution

DNL Differential Non-Linearity

d-SiPM Digital Silicon Photomultiplier

DCR Dark Count Rate

DOI Depth Of Interaction

ENOB Effective Number Of Bits

FF Fill Factor

FLIM Fluorescence-Lifetime Imaging Microscopy

FOM Figure Of Merit

FOV Field Of View

FPGA Field-Programmable Gate Array

FSM Finite State Machine

HEP High Energy Physics

INL Integral Non-Linearity

LiDAR Light Detection and Ranging

LOR Line Of Response

LSB Less Significant Bit

LSO Lutetium Oxyorthosilicate

LYSO Lutetium Yttrium Orthosilicate

MRI Magnetic Resonance Imaging

NaI Sodium Iodide

PDE Photon Detection Efficiency

PDP Photon Detection Probability

PEB Premature Edge Breakdown

PET Positron Emission Tomography

PLL Phase Locked Loop

PMT Photomultiplier Tube

PQARC Passive Quenching and Active Recharge Circuit

PQRC Passive Quenching and Recharge Circuit

PVT Process, Voltage and Temperature

PWO Lead Tungsate

QE Quantum Efficiency

QRC Quenching and Recharge Circuit

SiPM Silicon Photomultiplier

SNR Signal to Noise Ratio

SPAD Single-Photon Avalanche Diode

SRAM Static Random Access Memory

SRH Shockley-Read-Hall

STI Shallow Trench Isolation

TCAD Technology Computer-Aided Design

TDC Time to Digital Converter

ToF Time of Flight

VCRO Voltage Controlled Ring Oscillator

VLSI Very Large-Scale Integration

XCI Cross-Coupled Inverters

XCP Cross-Coupled PMOS

ZnS Zinc Sulfide

ABSTRACT

This thesis presents a contribution to the design of single-photon detectors for medical imaging. Specifically, the focus has been on the development of a pixel capable of single-photon counting in CMOS technology, and the associated sensor thereof. These sensors can work under low light conditions and provide timing information to determine the time-stamp of the incoming photons. For instance, this is particularly attractive for applications that rely either on time-of-flight measurements or on exponential decay determination of the light source, like positron emission tomography or fluorescence-lifetime imaging, respectively. This thesis proposes the study of the pixel architecture to optimize its performance in terms of sensitivity, linearity and signal to noise ratio.

The design of the pixel has followed a bottom-up approach, taking care of the smallest building block and studying how the different architecture choices affect performance. Among the various building blocks needed, special emphasis has been placed on the following:

- the Single-Photon Avalanche Diode (SPAD), a photodiode able to detect photons one by one;
- the front-end circuitry of this diode, commonly called quenching and recharge circuit;
- the Time-to-Digital Converter (TDC), which determines the timing performance of the pixel.

The proposed architectural exploration provides a comprehensive insight into the design space of the pixel, allowing to determine the optimum design points in terms of sensor sensitivity, linearity or signal to noise ratio, thus helping designers to navigate through non-straightforward trade-offs.

The proposed TDC is based on a voltage-controlled ring oscillator, since this architecture provides moderate time resolutions while keeping the footprint, the power, and conversion time relatively small. Two pseudo-differential delay stages have been studied, one with cross-coupled PMOS transistors and the other with cross-coupled inverters. Analytical studies and simulations have shown that cross-coupled inverters are the most appropriate to implement the TDC because they achieve better time resolution with smaller energy per conversion than cross-coupled PMOS transistor stages.

A $1.3 \times 1.3\,\mathrm{mm}^2$ pixel has been implemented in an 110 nm CMOS image sensor technology, to have the benefits of sub-micron technologies along with the cleanliness of CMOS image sensor technologies. The fabricated chips have been used to characterize the single-photon avalanche diodes. The results agree with expectations: a maximum photon detection probability of 46 % and a median dark count rate of $0.4\,\mathrm{Hz}/\mu\mathrm{m}^2$ with an excess voltage of 3 V. Furthermore, the characterization of the TDC shows that the time resolution is below 100 ps, which agrees with post-layout simulations. The differential non-linearity is ± 0.4 LSB, and the integral non-linearity is ± 6.1 LSB.

Photoemission occurs during characterization - an indication that the avalanches are not quenched properly. The cause of this has been identified to be in the design of the SPAD and the quenching circuit. SPADs are sensitive devices which maximum reverse current must be well defined and limited by the quenching circuit, otherwise unwanted effects like excessive cross-talk, noise, and power consumption may happen. Although this issue limits the operation of the implemented pixel, the information obtained during the characterization will help to avoid mistakes in future implementations.

CHAPTER 1

INTRODUCTION TO SIPM SENSORS FOR MEDICAL APPLICATIONS

From the beginning of medicine, there has always been interest in studying the inside of the human body. It has not been until the development of the first medical imaging techniques that it has been possible to satisfy such curiosity without invasive techniques, like surgery and catheterisation. One must go back to the end of 1800 to find the first attempt to realize medical imaging. This can be attributed to W. Röntgen and his studies on X-rays, thanks to which, he was awarded a Nobel prize. As many other scientists in that era, W. Röntgen refused to patent any of his inventions, because he understood the potential of that technology and considered that the technique would be improved to the extent that it was more accessible. This altruistic gesture and the natural curiosity of humankind lead to the first X-ray machines composed of high voltage batteries, as a power supply; a photographic plate, as a detector; and a X-ray tube, a vacuum tube evolved from the Crookes tube initially used by W. Röntgen. Despite of its simplicity, the images obtained from these rudimentary machines were able to provide valuable information about bone fracture, malformations, etc. However, since the consequences of exposure of living tissue to ionizing radiation were unknown at that time, those machines lacked shielding to confine radiation. The first X-ray operators suffer from overexposure and, eventually, died from its effects. It took long time, almost a century, and several lives to realize that no dose of ionizing radiation is free from risk. As time advance, technology allowed the development of more sensitive systems, like X-ray computed tomography (CT) scans that provide three-dimensional information of the inside of the body.

The understanding of other physical phenomena and the continuous im-

provement of technology have resulted in the development of other medical imaging techniques such as: magnetic resonance imaging (MRI), based on nuclear magnetic resonance of hydrogen atoms inside a magnetic field; ultrasound imaging, based on the transmission and refection of sound wave through the body to determine its structure; or positron emission tomography (PET), based on the detection of gamma photon pairs from the decay of a radio-tracer. Comparing PET with the other techniques, its main peculiarity is that it provides not anatomical but functional information of the body.

This chapter provides a brief historical review of the most important milestones in electronics and medical imaging. It also gives the needed background to understand the requirements of time of flight (ToF)-PET scanners and its sensors. Furthermore, this chapter tries to build up a bridge between the application designers and the sensor designers in order to set a common dictionary. This will facilitate the communication and discussion among them.

1.1 Electronics and Medical Imaging: a Brief History

A disclaimer must be done here, the history of science, and humanity itself, can not be attributed to few individuals or milestones. It is full of serendipities, independent parallel findings and, mostly, work based on the contribution of scientists who preceded us. However, it is easier for us to reduce the history to a time line, dotting the important dates, big discoveries or inventions. It should not be forgotten that this is nothing more than a simplification. Take as an example the radio communication: across the world, several scientists and inventors, like G. Marconi, N. Tesla, O. Lodge and A. Popov among others, were able to develop a technology to transmit information in electromagnetic waves. Although their contribution is undeniable, all these names were part of something bigger, a society, that gave them a reason to communicate. Being aware of this and having done the disclaimer, it must be accepted that it is impossible to synthesize the history of electronics and medical imaging in the following paragraphs. If the reader is interested in digging a little bit more about this topic, he or she can refer to the bibliography.

Electronics and medical imaging have been bound from their beginning, since they share common ancestors, the Crookes tube and the vacuum tube. The Crookes tube, invented by W. Crookes around 1870, is composed of two electrodes (anode and cathode) placed at opposite ends of a low pressure gas filled tube. Under high bias voltage, the electrons gain enough energy to travel from anode to cathode. A fraction of these electron hits the glass wall behind the cathode and excite its electrons. When these excited electrons return to their ground state, they emit visible light, the so called cathode rays. The same device under higher bias voltage is able to emit X-rays through Bremsstrahlung

Figure 1.1: Hand with ring: one of the first radiographs taken by W. Röntgen (1872–1919) of his wife Anna Bertha Ludwig's hand (1872–1919). January 1895 [1].

process. That is, due to the deceleration of the electron when they are deflected by other charged particle. Biasing the Crookes tube under such condition and with the help of a photosensitive cardboard made of barium platino-cyanide, W. Röntgen started paving the way of medical imaging in late 1895, by capturing the radiograph of Anna Bertha Ludwig hand's as figure 1.1 illustrates [1, 5]. In 1904, J. A. Fleming also started paving the way of electronics using thermionic emission principle to develop a vacuum tube that works as a rectifier. Years later, in 1907, Lee de Forest presented its "audion", a tree-terminal vacuum tube able to amplify the input signal, laying the foundations of electronics [6].

The discoveries in mathematics, physics and engineering served as a substrate for the development of new medical imaging techniques. Concepts like, Fourier transform, nuclear magnetic resonance, piezoelectricity, X-ray emission, absorption and its effects on the living tissue were the result of the work of the scientific community as a whole during the last centuries. Meanwhile, the advances in electronics, like the development of the first solid state transistor by W. Shockley, J. Bardeen and W. Brattain in 1947 and the fabrication of the first integrated circuit by J. Kilby in 1959, meant a paradigm shift to a much more compact, cheaper, scalable and less power hungry systems. Thanks to these discoveries, and a few more, between 1940 and 1955, the first medical ultrasound scanners were reported by K. and F. Dussik [5], I. Edler [7] and J. Wild [8]. The next important milestones in the history of medical imaging are concentrated in the decade of 1970. Firstly, the development of the CT scanner

in 1972 [5]. These scanners uses X-rays to generate cross-section images of body structures based on their density. Secondly, the MRI was reported in 1973 by P. Lautertur [9]. A MRI scan provides the distribution of body structures with high content of hydrogen, such as water and fat, using the resonance of hydrogen atoms inside a high magnetic field. Since MRI is not base on ionizing radiation, it is consider safer than CT. Thirdly, in 1975, A group composed of M. Ter-Pogossian, M. Phelps, E. Hoffman, N. Mullani demonstrated the working principle of PET [10]. All these imaging techniques are not mean to substitute the previous ones, but to provide additional information about different body structures and their function.

Medical imaging has taken advantage of the benefits of technological scaling in microelectronics. The incremental improvement of fabrication processes and device miniaturization, not only allow the integration of smaller, less power hungry and computationally more powerful systems, but also improved detector sensitivity, spatial resolution and reduced noise. Additionally, mass production of microelectronics reduced the system cost, allowing the benefits of medical imaging techniques to be enjoyed by more people across the world.

1.2 Positron Emission Tomography

Unlike other medical imaging techniques, such as CT or MRI, that provides information about body structures, PET imaging provides functional information. This is achieved by mapping the presence of a radiopharmaceutical, used as a tracer, associated to the function that must be highlighted. This technique has proved to be useful to study the brain activity and cancer metastases. Besides, some studies show that it is possible to highlight infectious diseases [11]. Generally, fludeoxyglucose is the radiotracer used to highlight metabolic activity. This radiopharmaceutical is a glucose analog, composed of a radionuclei ^{18}F substituting one of the -OH groups of the glucose molecule. For this reason, the molecules of fludeoxyglucose are interpreted as glucose molecules by the cells and gathered inside them until the radionuclei decays. Consequently, the larger the metabolic activity, the larger the concentration of this molecule inside the cells. The radioactive decay of ^{18}F could be a β^+ process with 97 % of probability (1.1a) or a electron capture process with 3 % of probability (1.1b):

$$ {}^{18}_{9}F \rightarrow {}^{18}_{8}O + e^+ + \nu_e \tag{1.1a} $$

$$ {}^{18}_{9}F + e^- \rightarrow {}^{18}_{8}O + \nu_e \tag{1.1b} $$

where ${}^{18}_{9}F$ is the fluorine radionuclei with a half-life of 10 min [12]; ${}^{18}_{8}O$ is a stable oxygen isotope; e^+ is the positron, the anti-particle of the electron; ν_e is a neutrino; and e^- the electron. Electron capture process has no interest, because it does not emit a positron. Fortunately, this process only represent the 3 %. When the decay occurs, the resulting glucose is consumed by a mitochondrion.

The created positron will travel a mean free path of 1 mm [12], until it collides with an electron. The result of this collision is the annihilation of both particles and the production of a pair of gamma photons moving in opposite directions, each one with an energy of 511 keV [13]. Since the positron and the electron may have non-zero kinetic energy, the angle between the two trajectories is $180° \pm 0.25°$ FWHM.

A conventional PET scanner, like the one shown in figure 1.2-(a), is composed of a few detector rings. Each ring, in turn, is divided into several detector blocks, which are made up of a scintillator crystal and a photodetector. In a first approximation, when the pair of gamma photons is detected by the ring, a line of response (LOR) can be drawn between the hit detectors. Since there is no more information about the annihilation, the probability that the annihilation have occurred at a given point along the LOR is distributed uniformly. In these scanners, it is necessary to acquire several LORs to identify the points where the radiotracer is decaying to reconstruct the image. On the other hand, the ToF-PET scanner, as the one represented in figure 1.2-(b), is able to estimate the arrival time of gamma photons, which reduces the uncertainty on the annihilation point along the LOR. The spacial accuracy of these scanners is function of the temporal accuracy of the detector:

$$\sigma_{LOR} = \frac{\sigma_{CTR}}{2c} \tag{1.2}$$

where σ_{LOR} is the spacial accuracy in the LOR; σ_{CTR} is the coincidence time resolution (CTR), the temporal accuracy of the detector; and c is the speed of light. The performance of ToF-PET scanners is usually compared to conventional scanners by means of the CTR and the image quality improvement associated with it. The signal to noise ratio (SNR) improvement between ToF and conventional scanners can be calculated as [14]:

$$\frac{SNR_{PET-TOF}}{SNR_{PET}} = \sqrt{\frac{2\,D_{body}}{c\,\sigma_{CTR}}} \tag{1.3}$$

where $SNR_{PET-TOF}$ is the signal to noise ratio of the ToF-PET scanner, SNR_{PET} is the signal to noise ratio of the conventional PET scanner, D_{body} is the patient's body diameter and c is the speed of light. A reduction of CTR brings an improvement of image quality [15], a shorter acquisition time and a smaller radiation doses.

Notwithstanding the above, there are some non-idealities that have not being taken into account and deteriorate the image quality. Firstly, as figure 1.3-(a) illustrates, random coincidences are produced when two gamma photons from different annihilations are classified as if they had the same origin. This usually happen when some annihilations occur outside the gantry, or field of view (FOV), of the scanner or when the direction of the gamma photon travels away from the FOV. The most immediate solution to solve the problems of a limited FOV, it is to increase the number of detector rings as in [16]. Despite

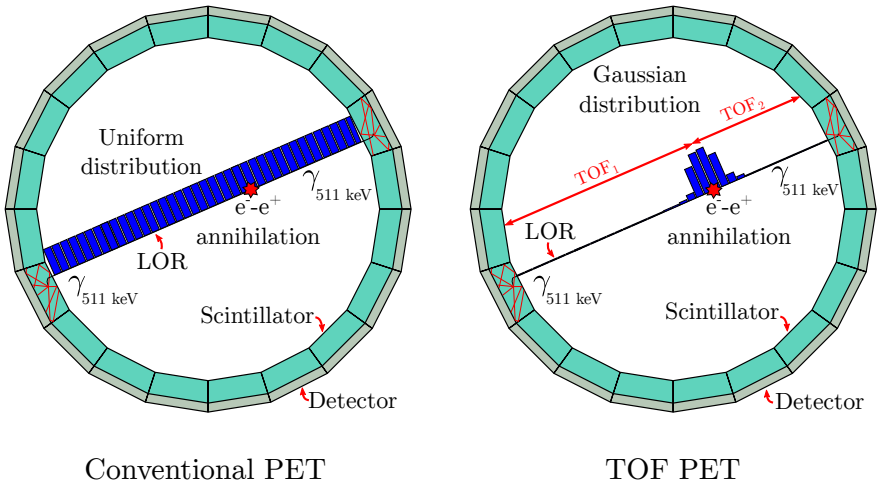


Figure 1.2: Positron emission tomography (PET) working principle: (a) conventional scanner and (b) time of flight (TOF) scanner.

the improvements in image quality and acquisition time, this is prohibitive on many occasions due to the high cost of the components, mainly scintillator crystals [17]. Secondly, gamma photons can also suffer several scattering along their trajectory. As figure 1.3-(b,c) represents, this can happen inside the patient's body or the scintillator. This kind of interaction between gamma photons and matter, changes the direction and reduces the energy of the gamma photon. Generally, in order to reduce the impact of this non-ideality, only events with an energy close to 511 keV are considered during the image reconstruction, the others are discarded. Also, to reduce the probability that the gamma photons escapes from scintillator, high density crystal, like lutetium oxyorthosilicate (LSO), bismuth germanate (BGO) or lead tungsate (PWO), are preferred over others less dense, like cesium iodide (CsI) or sodium iodide (NaI).

1.3 Scintillation Phenomena and Light Propagation

Since the discovery of the scintillation phenomena of zinc sulfide (ZnS) in 1903 by W. Crookes [18], a lot of effort has been done by the community to understand the physics behind and to develop new compounds [2]. It is indispensable for the electronic designer to understand the most basic physical phenomena occurring during a scintillation, like the transformation of gamma photons into optical photons, light propagation and extraction from the scintillator to the sensor. This is important because the scintillator is a key part of the sensor interface, since it transforms the ionizing energy of the gamma photons into

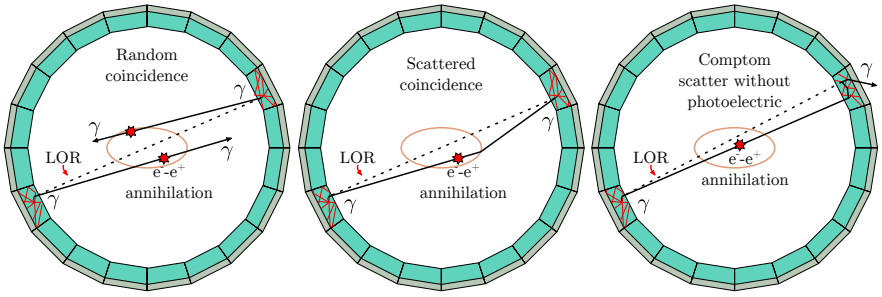


Figure 1.3: PET non-idealities: (a) random coincidence, (b) scatter coincidence and (c) Compton scatter without photoelectric absorption.

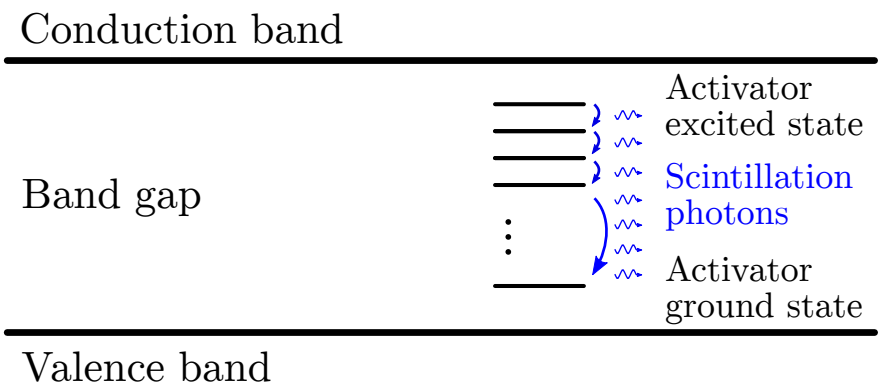


Figure 1.4: Band diagram of a scintillator crystal.

visible photons that can be detected by silicon photodiodes. To properly design the sensor, it is important to understand the way this transformation is done, its magnitude, dynamics and spatial distribution. It is also necessary to be aware of secondary phenomena that may have the key for producing prompt photons which will bring a CTR improvement in a near future, like Čerenkov radiation or hot intraband luminescence [19].

For PET scanners, inorganic scintillators have been used because of their larger light yield (photons/MeV), radiation hardness and fast response, compared to organic or plastic scintillators. From the point of view of light generation, these crystals can be classified into few categories, depending on which type of recombination center is present in their crystalline lattice: doping impurity or activator, like cerium/calcium doping in LSO and lutetium yttrium orthosilicate (LYSO); self-trapped excitons, like CsI; luminescence ion, like Bi^{+3} in BGO; and may others [2, 20].

The emission of optical photons in an inorganic scintillator can be explained

with its band diagram, as the one shown in figure 1.4. Scintillator crystals can be considered as insulators, since its band gap is high enough to prevent free electrons in the conduction band at room temperature, in the order of 5 eV [21]. This means that the electrons in the valence band are bound to the crystal lattice, while the conduction band is empty. In scintillators with doping impurities as recombination centers, like Ce^{+3}, Eu^{+2} or Tl^{+} [2], the impurities add several discrete energy levels in the band gap, without changing the valence or the conduction band. In these recombination centers, the excited electron can be here recombined with an hole emitting a large number of photons in the visible spectrum [20, 21]. On the contrary, some scintillators, like BGO, does not need a doping impurity, because its crystal lattice include a luminescence ion intrinsically. In the case of BGO, the Bi^{+3} ion is the recombination center of the crystal lattice where electrons returns to their ground state [20].

All in all, a scintillation event can be summarized as follows: an ionizing particle, like a gamma photon, provides enough energy to an electron in the valence band and generates a free electron-hole charge. The hole travels to a recombination center, since the ionization energy of crystal lattice is higher than the ionization energy of recombination center. The electron also travels to find the ionized recombination center. At this point, the recombination center becomes neutral and the electron starts its de-excitation, emitting thousands of photons in the ultraviolet and visible spectrum [20]. This emission is isotropic, that is, in any direction, and its dynamics can be described with one or several exponential decays [18]:

$$N = A\,\exp\left(-\frac{t}{\tau_f}\right) + B\,\exp\left(-\frac{t}{\tau_s}\right) \tag{1.4}$$

where N is the total number of emitted photons at time t, τ_f is the fast time constant, τ_s is the slow time constant, A and B are relative magnitudes of light yield that depends on the scintillator. LSO and LYSO are commonly modeled with one time constant of 40 ns, while PWO is modeled with two time constant: $\tau_f = 10\,\text{ns}$ and $\tau_s = 30\,\text{ns}$ [22]. A summary of some scintillator properties is given in table 1.1. In spite of the above, several interactions may occur between a gamma photon and the crystal electrons. Mostly, these interactions can be divided into Compton scattering and photoelectric absorption [23]. In a photoelectric absorption, the whole energy of the gamma photon is transfer to the electron. While, in a Compton scattering, the gamma photon transfer a fraction of its energy to the electron, generating a free electron-hole pair. The new trajectory of the gamma photon changes an angle proportional to the amount of transferred energy [24]. Since the gamma photon can suffer several Compton scatters before undergo in photoelectric interaction, the event energy is distributed across the crystal. When this happens, it is very difficult to estimate the entrance point of the gamma photon and, therefore, to reconstruct the LOR, as figure 1.3-(c) illustrates. To overcome this limitation, new architectures of scanners are being proposed. For instance, in [23], the LOR is reconstructed

Table 1.1: Properties of common scintillator crystal used in medical and high energy physic applications [2, 3, 4].

	NaI:Tl	CsI:Tl	BGO	PWO	LSO:Ce
Density (g/cm^3)	3.67	4.51	7.13	8.3	7.4
Light yield (photons/MeV)	38000	68000	8200	300	27000
Energy resolution (%)	7.4	7.9	10.8	75	9.1
Time constant (ns)	245	1220	300	10 30	40
Emission peak (nm)	410	550	480	420	420
Refractive index (-)	1.85	1.79	2.15	2.2	1.82

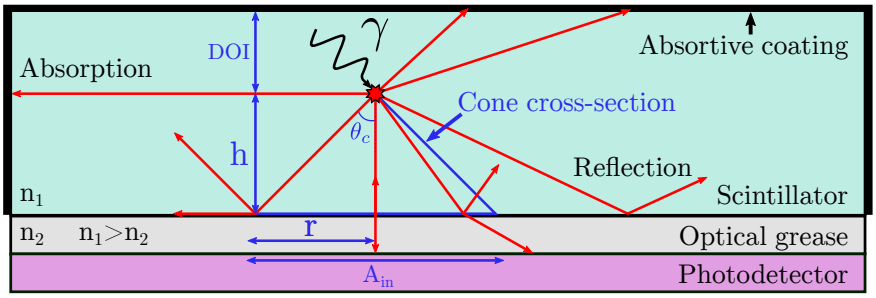


Figure 1.5: Photoelectric event and light propagation phenomena in a detector block composed of a scintillator crystal, an absorptive coating, an optical coupling medium and a photodetector. If $n_1 > n_2$, total internal reflection occurs and the light profile over the sensor can be simplified to a cone shape with radius r and height h.

taking into account also the information from Compton interactions, thanks to a multi-layer structure.

Once the optical photons have been created inside the crystal, the transmission, refraction and reflection become the most important phenomena for the designer, since they shape the light distribution over the photodetector. Monolithic crystal are preferred over pixelated crystals for several reasons. First of all, as figure 1.5 illustrates, with monolithic crystals, the spatial distribution of light during the absorption of the gamma photon can be sampled and studied. This allows to reconstruct and estimate the position where optical photons have been generated [25, 26]. Generally, this is done by calculating depth of interaction (DOI). In order to increase the light extraction, the crystals usually have a large surface coupled to the photodetector and small thickness, i.e., $10 \times 50 \times 50\,\text{mm}^3$ [27].

Additionally, it is evident that the number of reflections and absorptions inside the crystal must be minimized to maintain the spatial distribution of light. Otherwise, the information contained in the light distribution is loose.

In this sense, scintillators are transparent, with transmission coefficient close to 80 % [22], respect to their emission spectrum. However, this may not be the case for Čerenkov photons whose spectrum is close to the ultraviolet. For these wavelengths the transmission coefficient is small and Čerenkov photons are mostly absorbed by the crystal. As figure 1.5 represents, the interface between the crystal and the photodetector is composed of optical coupling, generally optical grease. If the refractive index of the scintillator is higher than the refractive index of the optical coupling, total internal reflection occurs. This phenomena imposes a critical angle, from which each incident photon is reflected. This critical angle is given by the Snell's law:

$$\theta_c = \arcsin\left(\frac{n_1}{n_2}\right) \tag{1.5}$$

where θ_c is the critical angle, n_1 is the refractive index of the scintillator and n_2 is the refractive index of the optical coupling. This angle limits the amount of photons reaching the sensor surface. For example, the refractive index of LSO is 1.82 and, typically, optical grease has a refractive index of 1.5, this leads to a critical angle of 56.44°. In order to improve light extraction, some methods proposed the use of photonic crystals at the surface of the crystal. They act as a matching impedance between the scintillator and the coupling medium [28, 29, 30]. The critical angle provides useful information about the light shape over the sensor and its magnitude. In a conventional set-up, like the one of figure 1.5, the amount of optical photons hitting the photodetector can be estimated. The worst case scenario is given by photoelectric events close to the detector surface, because it brightness can saturate the photodetector if it is not properly designed. Considering that photoemission is isotropic, the solid angle associated to the critical angle from (1.5) is:

$$\Omega = 4\pi \sin^2\left(\frac{\theta_c}{2}\right) \tag{1.6}$$

where Ω is the solid angle of the cone with its apex angle of $2\theta_c$ and its basis at the detector surface, see figure 1.5. The amount of photons leaving the crystal, without considering black reflections, can be estimated by the fraction of the generated photons during the scintillation and the solid angle from (1.6):

$$P_{out} = P_{total}\ \frac{\Omega}{4\pi} = LY\ E\ \sin^2\left(\frac{\theta_c}{2}\right) \tag{1.7}$$

where P_{out} is the number of photons leaving the crystal, P_{total} corresponds to the total number of photons generated during the scintillation, Ω is the solid angle between the scintillator and the optical coupling, 4π is the total number of steradians of a sphere, LY is the light yield of the scintillator, E is the energy of the gamma photon. From expression (1.7), if the thickness of the optical grease is neglected, the maximum density of optical photons per area can be calculated as:

$$H_q = \frac{P_{out}}{A_{in}} = \frac{LY\ E}{\pi}\left[\frac{\sin(\theta_c/2)}{\tan(\theta_c)\ h}\right]^2 \tag{1.8}$$

where H_q is the cumulative optical photons per area, A_{in} is the incidence area of the sensor, h is the distance between the recombination center and the crystal surface. Considering an interface between LSO crystal ($LY = 27\,000$ photons/MeV and $n = 1.82$) and optical grease ($n = 1.5$); a photoelectric interaction of 511 keV with a distance $h = 1$ mm, the critical angle is $\theta_c = 56.44°$ and the maximum density of optical photons per area is:

$$H_q = \frac{27000 \cdot 0.511}{\pi}\left[\frac{\sin(56.44/2)}{\tan(56.44)}\right]^2 \approx 450\,\text{photons/mm}^2 \tag{1.9}$$

This upper bound value of $H_q = 450$ photons/mm^2 only takes into account the forward emission. However, it provides an idea of the maximum magnitude of input light and shows that the sensor will work in single-photon regime, even when the gamma photon deposits all its energy close to the sensor.

1.4 Photodetector

Once the optical photons have left the crystal, they must be detected and translated into electrical signals that can be processed. To do so, the first PET scanners rely on photomultiplier tube (PMT)s as a photodetectors. The PMT have their origin in the vacuum tubes developed by H. Iams and B. Salzberg in 1935 [31]. The figure 1.6 represents the working principle of conventional PMT. An incident photon with enough energy hit the photocathode and, thanks to the photoelectric effect, a free electron is produced. Next, this electron is accelerated toward the first dynode, due to the high electric field between the photocathode and the dynode. Then, the collision with the first dynode generates more electrons, that at the same time are attracted to the second dynodes. This multiplication process finishes at the anode where the signal is sampled by the front-end electronics. Thanks to their high gain, linearity, robustness, radiation hardness, fast response and low noise, these devices were the best option for a broad range of applications, not only single-photon detection, since their adoption in the early 1940s. Usually, the main drawbacks of these devices are associated with their high bias voltage, power consumption and non-compatibility with magnetic fields.

During the last decades, the silicon photomultiplier (SiPM) has rose as a solid state alternative to substitute PMTs in many applications. As figure 1.7 represents, SiPM are composed of an array of SPADs with quenching circuitry connected in parallel and biased on the Geiger mode, i.e., reversed biased above the breakdown voltage. Broadly speaking, the working principle of these devices can be summarize as follows: an incident photon with enough energy is absorbed at the diode depletion region and, thanks to the photoelectric effect,

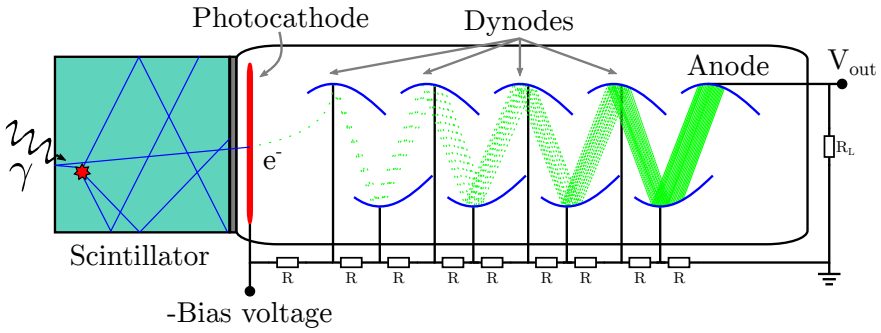


Figure 1.6: Photomultiplier tube working principle. The photoelectric effect generates a electron at the photocathode. The electrons are drifted by the electric field the the dynodes, where more electrons are emitted.

a free electron-hole pair is produced. These free charge carriers are accelerated by the high electric field in the depletion region. The collisions of these charge carriers with the crystal lattice release more electron-hole pairs in a process called impact ionization [32, 33]. This sharp current rise, also called avalanche, is detected by the front-end electronics. Another function of the front-end electronics is to limit the amount of current flowing through the SPAD, quench it and reset the device to its original bias condition, ready for another detection. Compared to PMT, the SiPM has a smaller power consumption, bias voltage and it is compatible with magnetic fields and more compact, allowing higher spatial resolution. The main drawbacks of these devices are their high noise and low radiation hardness.

In figure 1.7-(a) the schematic of an analog SiPM is shown, each one of the micro-cells of this device is composed of a SPAD and a passive quenching resistor. The output of an analog SiPM provides a current proportional to the number of detected photons. In order to optimize the quantum efficiency (QE), reduce noise and increase fill factor, analog SiPMs are usually fabricated into custom process. On the contrary, digital SiPMs are fabricated with CMOS or CMOS image sensor (CIS) technologies. As figure 1.7-(b) represents, the main advantage of this approach is the flexibility that the designer have to implement processing logic close to the focal plane of the SiPM, i.e. to determine the photons arrival time with a TDC. This flexibility comes at the expense of a degradation of QE, noise and fill factor. Before evaluating and comparing the performance of the state of the art, several parameters and figure of merit (FOM) must be defined briefly:

- Photon detection probability (PDP) is the ratio between the detected photons and the incident photons without considering losses due to dead area. It depends on the technology, incident wavelength, the SPAD structure and the bias voltage.

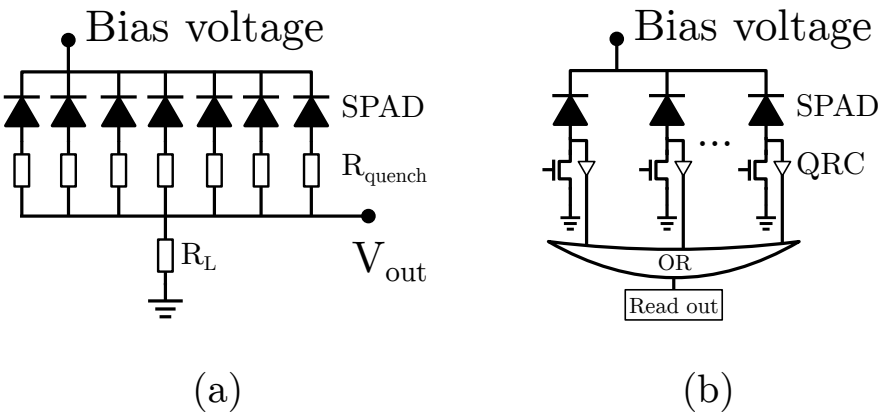


Figure 1.7: (a) Analog silicon photomultiplier (SiPM) and (b) digital SiPM with common read out.

- Fill factor (FF) is the ratio between the photo-sensitive area and the dead area of the device, micro-cell, pixel or chip. The device fill factor is limited by the technology, since every SPAD requires a surrounding structure, guard ring, to confine the high electric fields. Without a proper guard ring the functionality of the SPADs is not guaranteed.

- Dark count rate (DCR) is the number of avalanches counted when the device is in complete darkness. This source of noise is uncorrelated with other sources and it is function of temperature and bias voltage.

- Cross-talk is the probability that one SPAD provokes avalanches in neighboring SPADs. A cross-talk event can be trigger by diffused charge carriers or photon generated by avalanche photo-emission.

- After-pulsing is the probability that a secondary avalanche is triggered due to the delayed release of charge carriers from deep level traps, like impurities and defects in the silicon lattice. These deep level traps are filled during a previous avalanches, and released sometime after that.

- Coincidence time resolution (CTR) is the time accuracy of two detectors working in coincidence. In a PET scanner, it takes into account the contribution of all the sources of uncertainty or jitter. From the system clock jitter to the SPAD and TDC jitter.

The state of the art of SiPMs for ToF-PET is summarized in table 1.2. It can be observed that custom technologies provide better performance in terms of PDP and DCR. However, they require additional circuit to estimate the photons arrival time. Despite having worst performance, digital technologies allow the

Table 1.2: State of the art of Silicon Photomultiplier (SiPM).

	[37]	[38, 39]	[40, 41]	[42, 43]	[44]	[45, 46]
Architecture	Analog	Analog	Analog	Digital	Digital	Digital
Technology (nm)	Custom	Custom	Custom	180 Custom	130 CIS	350 HV
SiPM pitch (mm)	3	3	4	4	0.59	0.8
SPAD size (μm^2)	2500	1225	2500	3780	210	1200
SPAD/SiPM	3531	5676	6400	3200	720	416
PDP (%)	68	51	76	54	45	32
FF (%)	74	75	42	74	43	57
DCR ($Hz/\mu m^2$)	-	0.05	0.2	0.15	65	50
Cross-talk (%)	7	8	-	7	2	1.8
After-pulsing (%)	-	0.25	-	0.1	0	-
CTR (ps)	180*	220*	200*	325**	400*	400*

* Laboratory set-up; ** PET-TOF scanner.

integration of on-chip processing logic, like event detection, TDC and communication logic. Nevertheless, the area occupied by these logic has a negative impact on the FF. Also, the DCR of a digital SiPM is two order of magnitude higher than that of their analog counterpart. Apart from ToF-PET, some other applications, like fluorescence-lifetime imaging microscopy (FLIM) [34] and light detection and ranging (LiDAR) [35], also benefit from on chip processing logic. However, it seems evident that the current trend is to implement the designs into 3D-stacked technologies [35, 36]. Unfortunately, nowadays this kind of processes are not widely available. To the best of the author knowledge, the honors of bringing together the best of analog and digital SiPMs is not yet claimed.

1.5 Case Study and Desired Features (Thesis Goals)

As it has been shown in previous sections, PET scanners can improve their image quality via SNR improvement applying ToF. The technology scaling that has been pushing the microelectronics industry to its limits year after year has improved the working frequency, power consumption and occupation area of consumer electronics. Almost every circuits in a digital SiPM benefits from this trend, specially the TDC in terms of temporal resolution. However, technology scaling does not always improves SiPM performance. The SPADs are the most affected devices when they are implemented in small technology nodes. This is because they suffer from low yield, due to the high doping concentration and non-spatial uniformity of the doping layers used to implement the device. To slightly alleviate this, CIS technologies are good alternatives. Although they have not been completely optimized for SPADs, these technologies contain less impurities and defects than standard CMOS. Also, they offer some properties attractive for SiPMs, like anti-reflective coating. With this in mind, a 110 nm CIS

technology has been chosen to design the digital SiPM for ToF-PET. 3D-stacked technologies have not been tested because of their high cost and because there is still room at the bottom to exploit the performance of SiPM in CMOS and CIS technologies.

The main goal of this thesis is to design a digital SiPM able to perform three main functions: to detect and count photons, to measure their arrival time and to process the scintillation event. Single-photon detection can be performed in CIS technologies with SPADs. However, the integration of such full custom devices is not trivial, because it requires a careful design and a deep knowledge of the process. The technology also allows the integration of active quenching and recharge circuit (AQRC) and digital logic to count the pulses generated by the avalanches. Moreover, time measurements can be done with a TDC. Since time resolution depend on the TDC architecture, it is necessary to study which architecture fits better this application. Specifically, a small area footprint is required to reduce dead area and increase fill factor. Also, a small conversion time is preferable to reduce temporal blind spots. Furthermore, power consumption should be also low to reduce the variations associated to thermal gradients. Additionally, time resolution should be small enough for not to be the dominant source of uncertainty. Also, an event processing finite state machine (FSM) must be flexible enough to study different trigger logic and threshold alternatives. For the proposed prototype, such flexibility can be achieved by reconfigurable logic, preferably a field-programmable gate array (FPGA).

Once the three main blocks and the rest of auxiliary logic are designed, the next step is to start the architectural exploration of the SiPM. This optimization process relies on the fact that each photon carries valuable information about its origin, therefore, photon detection and counting must be maximize. The architectural exploration provides an insight into the design space of these sensors, and helps to identify optimum points. This information helps the designer to justify their election and make the design more straightforward. The final step consist in the experimental characterization of the designed blocks and verification of the design approach.

1.6 Thesis Organization

Chapter 2 explains the design steps of two full custom SPADs and its AQRC in a CIS technology. In chapter 3, some TDC architectures are discussed and the design of a TDC based on VCRO is described. The architectural exploration of the SiPM is carried out in chapter 4 based on analytical methods. In chapter 5, the characterization of the SiPM is carried out. Chapter 6 discuss the main results and limitations of the design. Based on the experience gained during the design and characterization of the first prototype, the future work is described. Finally, chapter 7 provides the conclusions of this thesis.

CHAPTER 2

SPAD AND FRONT-END ELECTRONICS IN CMOS TECHNOLOGIES

SPADs are usually full custom devices, i. e. foundries do not provide them as a library blocks in standard CMOS technologies, not even in CIS technologies. Sometimes this means to break few of their design rules. Since these diodes are biased above the breakdown voltage of its sensing junction, special care must be taken during design to prevent disturbances to the surrounding logic. A common way to plan the design of complex circuits is to divide them into different levels of abstraction. At architectural level, the important metrics for each application need to be extracted, namely power consumption, bandwidth, QE, etc. At circuit level, the limitations of technology can set new boundaries to our study. At physical level, deviations caused by mismatch and process, voltage and temperature (PVT) variations must be considered to reduce its impact in the final implementation. In analog circuits, this planning is especially needed to guarantee signal integrity, since all transformations add distortion and noise. Despite SPADs are designed as full custom analog cells, their working principle allows designers to digitize their internal state: 'avalanche' and 'no-avalanche'. This is especially useful while working in single-photon regime, because this relates each digital pulse with an incident photon, if an ideal noiseless device is considered. Therefore, distortion would not affect, in principle, in the single-photon regime. In real devices, a fraction of the digital pulses are associated to noise. These false counts come from avalanches triggered by free charge carriers, generated by Shockley-Read-Hall (SRH) effect or band-to-band tunneling. Apart from the device itself, the design of its front-end is also

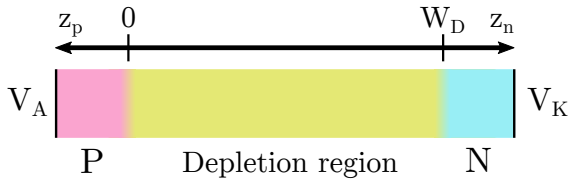


Figure 2.1: Cross-section of the p-n junction.

critical. Permanent damage may occur to the SPAD if the avalanches are not properly quenched. This is because avalanche current can reach magnitudes on the range of milliamperes, leading to thermal damage and, in the long term, electromigration.

This chapter is focused on the description of SPAD structures with low DCR and high PDP. The considered devices operates at 400–500nm range to match the output wavelength of LSO scintillator. Section 2.1 describes the working principle of an ideal SPAD. In section 2.2, the most used structures are presented and their main trade-offs are described. Section 2.4 describes the main characteristics of quenching and recharge circuits.

2.1 SPAD Working Principle

SPADs are diodes operating in reverse bias. Their bias voltage is beyond the breakdown voltage (V_{BD}). This is an unstable equilibrium point[1] at which any free charge carrier at the depletion region has a probability to trigger a self-sustained avalanche. Analytically, this conditions is achieved when the integral of the mean ionization rate for the charge carriers, $\overline{\alpha_{n,p}}$, in the depletion region width is larger than 1 [33],

$$1 < \int_0^{W_D} \overline{\alpha_{n,p}} dz \tag{2.1}$$

where W_D is the depletion region width as figure 2.1 illustrates. The integral in expression (2.1) is the so called avalanche condition. It states that the average distance for a free charge carrier to generate more carriers via impact ionization must be smaller the depletion region width [47]. When the avalanche is triggered in a semiconductor with different ionization rates for electrons and

[1] Here the term unstable equilibrium point is used rather than metastable point, because metastability is usually defined in electronics as a point with non-valid output. Generally, in digital electronics, a signal is metastable when its value is different from the defined states. In binary logic this means a value different than '0' or '1'. However, SPADs are analog devices whose state is defined, even when they are biased beyond its breakdown voltage, their output can be determined between 'avalanche' and 'no-avalanche'.

holes, expression (2.1) becomes:

$$1 < \int_0^{W_D} \alpha_n \exp\left[-\int_z^{W_D} (\alpha_n - \alpha_p) dz'\right] dz \tag{2.2a}$$

$$1 < \int_0^{W_D} \alpha_p \exp\left[-\int_0^{z} (\alpha_p - \alpha_n) dz'\right] dz \tag{2.2b}$$

where α_n and α_p are the ionization rates for electrons and holes, respectively. Expression (2.2a) describes the avalanche condition when the avalanche is initiated by electrons and expression (2.2b) when it is triggered by holes. Since both charge carriers can initiate avalanche, these two equations are equivalent [33]. The breakdown voltage, V_{BD}, can be derived numerically from (2.1) and the ionization rate for one-side abrupt junction [33]:

$$V_{BD} = \frac{\epsilon_s \xi_m^2}{2qN} \tag{2.3}$$

where ϵ_s is the dielectric constant of the semiconductor, ξ_m is the maximum electric field in the depletion region, q is the charge of the electron and N is the impurity concentration of the low-doped layer. Although this expression is an approximation, it shows that the breakdown is inversely proportional to the layer doping, a behavior that is observed experimentally [48, 49].

The polarity and current sense criteria commonly used in the bibliography is shown in figure 2.2-(a). The reverse I-V characteristic of a SPAD is depicted in figure 2.2-(b). When the voltage between the cathode and the anode is above the breakdown voltage ($V_{KA} = V_{BD} + V_{ex}$, with $V_{ex} > 0$V), the condition for avalanches to be self-sustained is meet. Under such bias, also called Geiger mode, the electric field at the depletion region is so high that any free charge carrier can be accelerated and gain enough kinetic energy to free more electron-hole pairs in a process called impact ionization [32, 33]. This biasing point is illustrated by the dotted line in figure 2.2-(b).

The seeding of an avalanche by an incident photon is depicted in figure 2.2-(c). The seed growth leads to an exponential current increment, until it saturates, as a result of a local lowering in the electric field. Then, the avalanche starts propagating from the seed point to the rest of the active area as figure 2.2-(c) illustrates [50, 47]. Two mechanisms are involved in the horizontal propagation of the avalanche. The first is the growth of the original seed by transversal carrier transport with a maximum spreading speed [50]. The second one is produced by the emitted photons during the avalanche [51]. This photon-assisted propagation creates new seeds at different locations of the active area [52]. These two processes are of a stochastic nature since the position where the incident photon is absorbed and its spreading are purely random processes. Generally, SPADs are connected in series with an impedance element to limit the current through them. Figure 2.2-(a) represents the simplest circuit that limits the avalanche current, a passive quenching and recharge circuit (PQRC). The effect of

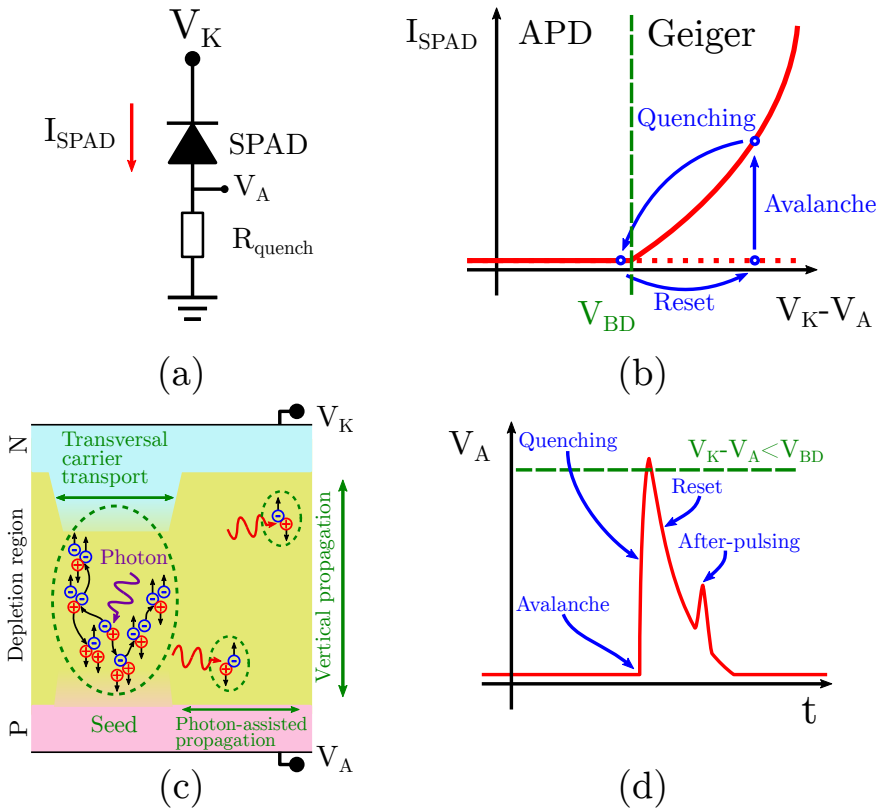


Figure 2.2: Single photon avalanche diode (SPAD) working principle: (a) SPAD with resistive quenching and recharge; (b) current as a function of voltage for a diode bias in the Geiger mode; (c) SPAD cross-section during avalanche with seeding point and photon-assisted propagation; and (d) transient voltage at the anode of the SPAD during avalanche.

this ballast resistor is an increase of the anode voltage as the current rise, hence a reduction of the cathode-anode voltage below the breakdown level, as it is represented by dashed lines in figure 2.2-(b, d). Note that the avalanche can only be quenched if the voltage across the SPAD terminals is below the breakdown voltage or, in other terms, if the current across the SPAD is below the minimum current that guarantee a self-sustained avalanche. Once quenched, the anode is discharged slowly down to zero volts, returning to the initial biasing point.

Since any free charge carrier can trigger an avalanche, these devices are sensitive to free charge carrier generated from photoelectric effect, Shockley-Read-Hall (SRH) effect and band-to-band tunneling. The former phenomena are the primary source of uncorrelated noise in SPADs. Another interesting phenomenon that is illustrated in figure 2.2-(d) is after-pulsing. It consist in

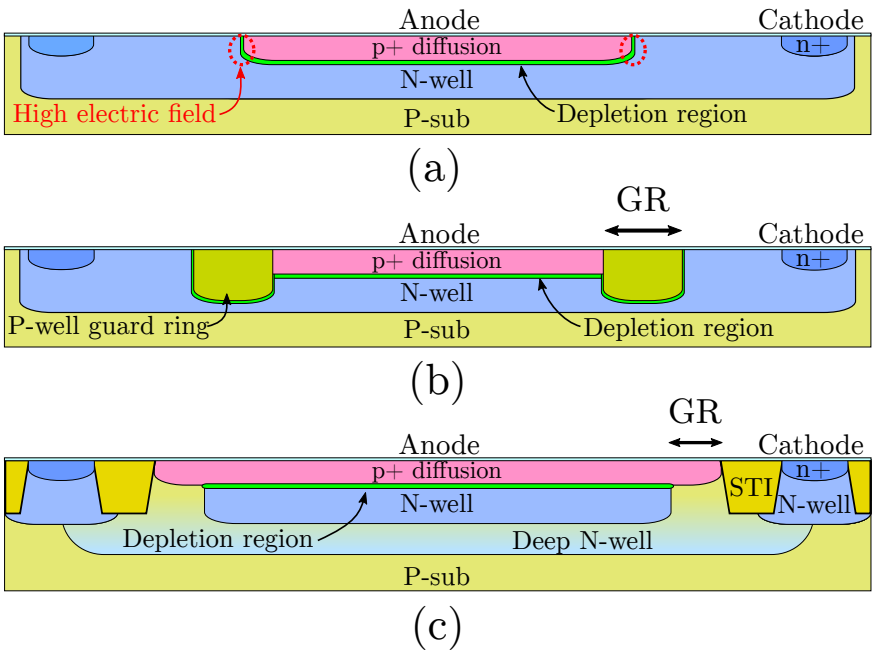


Figure 2.3: SPAD cross-section with superficial junction: (a) without guard ring, (b) with P-well guard ring and (c) virtual guard ring.

secondary triggers due to the release of charge carriers from deep level traps that were filled during the avalanche. It is therefore are a source of correlated noise. As expected, the larger the number of carriers crossing the junction, the larger the number of filled traps [53]. As it will be seen in the following section, designers can attenuate the impact of these noises with careful design.

2.2 SPAD Structures

All the characteristics of the SPAD are determined by its physical structure. PDP, DCR, jitter, etc. derive from a particular combination of physical layers. However, designers do not have often the necessary information to perform simulations with the help of technology computer-aided design (TCAD) tools, as the foundries are wary of providing sensitive data such as doping profiles or junction depths. The exploration of SPAD structures without the reliable data is a difficult task. Sometimes, several iterations are needed to infer and correct the errors.

Commonly, designers working with a CIS technology are limited to the layers that produce a p-n junction parallel to the silicon surface, i.e. the incident light is perpendicular to the diode active area. From all the possible combina-

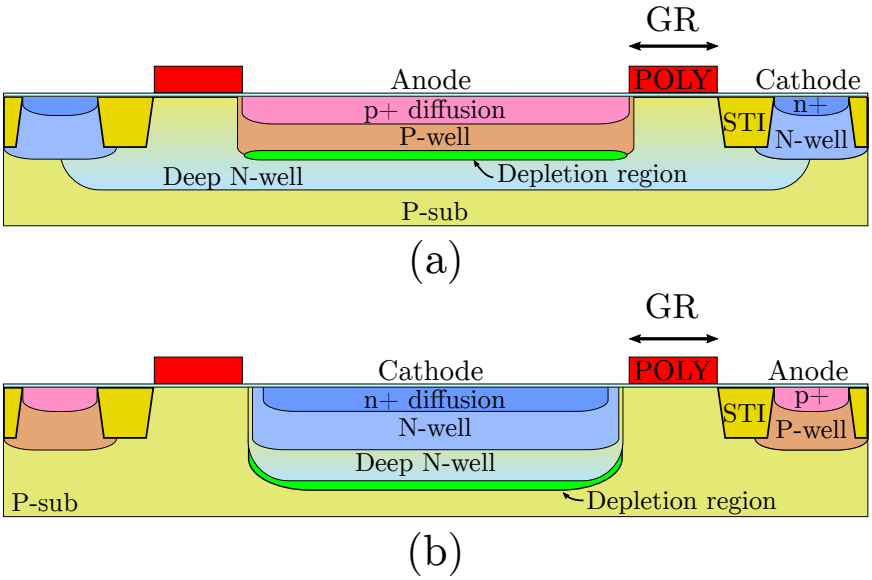


Figure 2.4: SPAD cross-section with deep junction: (a) red-sensitive structure and (b) near infrared sensitive structure.

tion of layers, the designers choose the SPAD structure that provides the best performance for their applications. A common way to classify these structures is using the junction depth, because a large number of SPAD parameters depends on it, directly or indirectly. For example, as in any other photodiode, the junction depth affects the wavelength at which the peak of sensitivity is found. In the following subsections, the benefits and drawbacks of SPADs implemented with superficial junctions [48, 54, 55, 56] (see figure 2.3) and deep junctions [48, 49, 54] (see figure 2.4) are discussed.

2.2.1 Noise

Noise in SPADs consists of spurious avalanches produced by charge carriers whose origin is different from photoelectric effect. Noise sources can be divided into uncorrelated and correlated. The dark count rate (DCR) groups together all uncorrelated noise sources. They are manly the result of thermal trap assisted carriers generated, by SRH, and carriers jumping from valence to conduction band, i.e. band-to-band tunneling. These phenomena are represented in figure 2.5 [57]. Although purely thermal carriers (no trap is involved) can also be generated by SRH, this noise source is less common in indirect band gap semiconductors, like silicon [46]. Generally, superficial p-n junction are composed of diffusion and well layers as indicated in figure 2.3. Since these structures have higher doping concentration, they are expected to suffer more DCR than

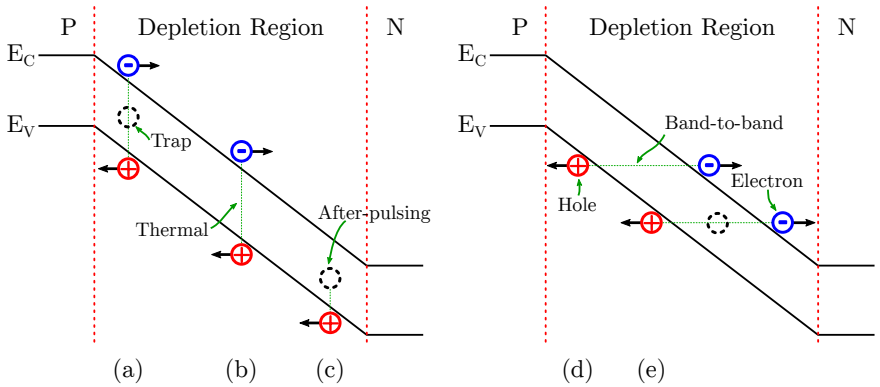


Figure 2.5: SPAD noise sources: (a) thermal trap assisted Shockley-Read-Hall generation, (b) thermal Shockley-Read-Hall generation, (c) carrier release from trap (after-pulsing), (d) band-to-band tunneling and (e) trap assisted band-to-band tunneling.

the deeper junctions composed of less doped layers, like those represented in figure 2.4. The main cause for this is the rise of band-to-band tunneling contribution due to a thinner depletion region [57].

On the other hand, correlated noise sources produce secondary avalanches after a primary avalanche has occurred. They can be classified in after-pulsing and cross-talk. An after-pulse is produced when a free charge carrier is captured by a deep-level trap during an avalanche and it is released some time later, see figure 2.5. These traps are impurities or defects in the silicon lattice. The after-pulsing probability can be calculated as [46]:

$$P_{ap}(t) = P_t \, P_{av} \, \frac{\exp(-t/\tau_t)}{\tau_t} \tag{2.4}$$

where $P_{ap}(t)$ is the after-pulsing probability at time t after the primary avalanche; P_t is the trap capture probability, it depends on the density of impurities in the lattice and the current flux during avalanche; P_{av} is the probability that a free charge carrier triggers an avalanche; and τ_t is the trap lifetime [46]. From the designer point of view, there are several ways to reduce after-pulsing. The first step is to choose a clean technology in order to reduce the number of impurities and defects that act as traps. Next, some SPAD structures suffer from more after-pulsing than others. For example, when shallow trench isolation (STI) are used as a guard ring, the interface states at silicon-SiO_2 boundary, act as traps [58, 59, 60]. Also, the maximum current during the avalanche can be limited to reduce the amount of trapped charges. Finally, the dead time of the SPAD can be adjusted to be larger than the trap lifetime. Doing so, a significant fraction of traps are discharged before the SPAD becomes sensitive again.

Cross-talk is the probability that one SPAD provokes avalanches in any of the neighboring SPADs as figure 2.6 illustrates. A cross-talk event can be

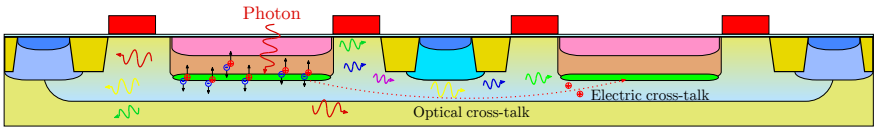


Figure 2.6: SPAD electrical and optical cross-talk.

initiated by diffused charge carriers or photons generated by avalanche photoemission [52, 61] traveling through the silicon bulk to the active area of a different SPAD [47, 46]. For these reasons, opaque and isolating guard rings are required. Unfortunately, sometimes this means to increase the guard ring area and, hence, to reduce the fill factor. The junction depth is especially relevant in this aspect, because superficial junctions are better confined by their guard rings, which are more effective as a physical barrier to surround the active area and to prevent particles generated during the avalanche to travel to other SPAD's active area. As reported in [54], SPADs with deep junction are less isolated between each other, because charge carriers are not so well confined.

2.2.2 Photon Detection Probability

The photon detection probability (PDP) is the ratio between the detected and incident photons for a normalized area. PDP can be calculated as:

$$PDP = QE(\lambda) \cdot P_{av}(V_{ex}) \tag{2.5}$$

where $QE(\lambda)$ is the quantum efficiency, i.e. the probability that a photon of a given wavelength, λ, generates a free electron-hole pair; and $P_{av}(V_{ex})$ is the probability that a free charge carrier triggers an avalanche, which is a function of the excess voltage (V_{ex}). Practically, this means that deeper junctions are more sensitive to longer wavelength, because the charge carriers generated by them are closer to the active area. This allows the designers to choose the structure that provides the PDP peak closer to the applications needs, selecting the more appropriate junction depth. This metric is more useful than photon detection efficiency (PDE) at this level, because it focuses on the performance of the technology and junction, not in the fill factor of the diode and the diode array.

As it is expected, SPADs with superficial junctions, like the ones of figure 2.3-(b,c), are preferred in applications where higher sensitivity to blue is needed [43, 62]. On contrary, the junctions represented in figure 2.4 are a common choice when high sensitivity to red or near-infrared is required [49, 62].

2.2.3 Jitter

Jitter in SPADs is related with the seeding point of the avalanche, its horizontal propagation and the active area of the device. As it was described in sec-

tion 2.1, the seeding point is the place where the incident photon creates the first electron-hole pair that initiates the avalanche. After that, the avalanche begins to spread to the rest of the active area and the current rises [50, 52]. Since the avalanche current depends directly on the avalanche propagation, its rise time is also affected by the stochastic nature of these processes. Therefore, if the avalanche is detected by the rising edge of the avalanche current, the jitter can be defined as the standard deviation, or FWHM, of the avalanche propagation time. It is important to notice that the size of the depletion region, both the total active area and the width of the depletion region itself (measured in a direction that is perpendicular to the junction), has an impact in the jitter. This is because the avalanche propagation time depends on the size of the depletion region. Statistical variations appear in the number of the emitted photons, so they have an impact in the jitter [52]. As of today, the smallest jitter reported corresponds to shallow diffusion/well junctions and small SPAD areas [48, 63, 56].

2.2.4 Premature Edge Breakdown

Premature edge breakdown (PEB) occurs in areas where the electric field is locally higher than in the rest of the junction. This high electric field entails a lower breakdown voltage than in the active area, because the breakdown condition represented in (2.1) is meet before in these regions. The areas more prone to this kind of failure are located where the doping profile has strong variations, like the periphery of diffusion, see figure 2.3-(a), or the corners of a square-shape SPAD [48]. As a result, the avalanches are triggered at the edges, rather than at the active area. This leads to a SPAD that is insensitive to incident light, since the active area never reaches a proper bias.

PEB can be avoided with a properly designed guard ring. Guard rings are implemented with layers that are less doped than the ones of the active area, surrounding it, in order to confine its high electric field. Generally, the used layers are P-wells or retrograde Deep N-wells, see figure 2.3 and 2.4. Although STI are effective to avoid PEB, the direct contact between active area and STI leads to a noisy device, due to the interface states at silicon-SiO_2 boundary [59]. PEB also affects diodes with sharp corners, for that reason, designers prefer circular shapes, or at least squared with round corners, rather than squared [48].

2.3 Implemented SPAD

In order to experimentally validate the technology, two different SPAD structures have been designed: a P-well/Deep N-well diode and P+/N-well diode. From now on PW3V3/DNW and PP/NW3V3, respectively. As figure 2.7-(a,b) illustrates, both structures share some common features. The deep N-well cathode can be shared among the neighboring SPADs to increase the fill factor. The poly-silicon layer is employed to form the P-substrate guard ring above the retrograde deep

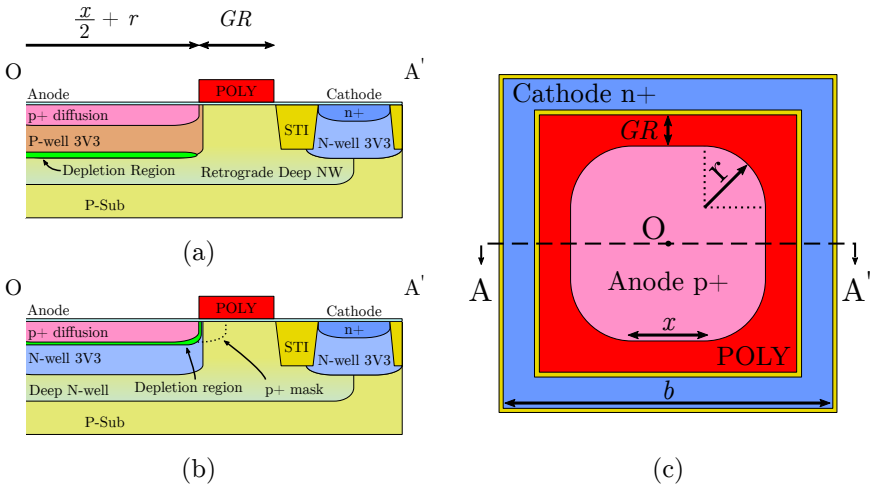


Figure 2.7: (a) PW3V3/DNW SPAD cross-section (b) PP/NW3V3 SPAD cross-section and (c) SPAD front view. Where r = 5µm, x = 5µm, GR = 1µm, b = 21.2µm, active area = 200µm^2.

N-well. Moreover, poly-silicon is used to prevent the formation of STI close to the depletion region. The technology provides both 1.2 V and 3.3 V wells, but it is expected that 3.3 V wells have lower and uniform doping than 1.2 V, because of technological scaling [64]. The effects of doping non-uniformity also affect superficial junction composed highly doped layers, like p+/n+ diffusion. It is worth mentioning that the PP/NW3V3 structure implemented will suffer from PEB. As seen in figure 2.7-(b), the p+ diffusion mask was extended to enclose the N-well as figure 2.3-(c) shows. However, this extended diffusion below the poly-silicon is not fabricated as in CMOS processes the gate self-aligns [33], therefore the poly-silicon acts as a stopping layer that prevents the formation of diffusion below it beyond a small overlapping. The effect of this error will be discussed in chapter 5.

It is expected then that PW3V3/DNW diode performs better in terms of DCR and PDP at long wavelength than PP/NW3V3. On the contrary, PP/NW3V3 structure may have smaller jitter, cross-talk and more PDP at short wavelength than PW3V3/DNW [60, 54, 48]. Comparing PW3V3/DNW against PP/NW3V3 structure, the main advantages should be less noise, more uniformity among devices and, therefore, higher yield.

Figure 2.7-(c) represents the layout of a SPAD parametric cell. A squared shape with round corners has higher fill factor than the circle-shaped and lower noise than square-shaped SPADs with straight corners. As for the dimensions, a small footprint will keep the DCR small, while the SPAD FF is close to 40 %:

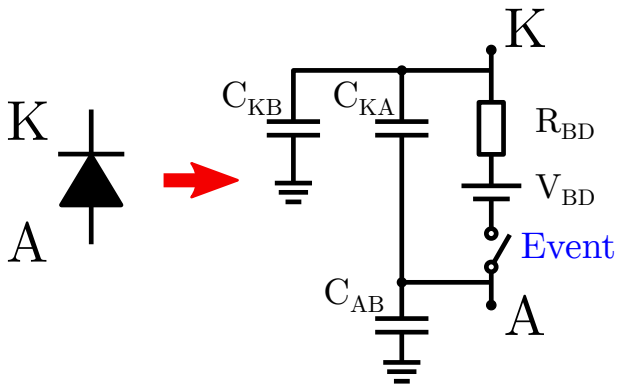


Figure 2.8: SPAD circuit model.

$$FF_{spad} = \frac{A_{aspad}}{A_{spad}} = \frac{x^2 + 4 \cdot r \cdot x + r^2\pi}{b^2} \tag{2.6}$$

where A_{aspad} is the active area of the SPAD, A_{spad} is the total area of the SPAD, x is the size of the straight segment of the SPAD size, r is the corner radius, $b = x + 2(GR + r + b')$ is the side of the complete structure and b' is the minimum size of n+ diffusion given by design rules of the technology, that in the example is $b' = 2.1\,\mu m$. Following the guidelines reported in [48, 54], these parameters can take values like $r = 5\,\mu m$ and $GR = 1\,\mu m$. Expression (2.6) can be solved for a SPAD fill factor around 40 % rendering $x \approx 5\,\mu m$.

2.4 Quenching and Recharge Circuit

Once the SPAD structure is selected and the dimensions defined, the next step is the design of the electronic front-end for the diode. The main component is the quenching and recharge circuit (QRC). The assembly of a SPAD and its QRC is usually known as a micro-cell. As it will be seen in chapter 5, the design of this circuit is critical, because the performance of the whole system depends on the work done at this stage. A bad design can mean anything from erratic operation and excessive power consumption to irrecoverable device damage. For this reason, it is very important to have a basic understanding of the SPAD working principle and to have an accurate model to perform simulations. It is relatively easy to find examples of both passive [65, 40, 66] and active [67, 68, 69, 70, 71] quenching and recharge circuit sizing.

The first step, before starting the design of the QRC, is the characterization of the SPAD and the extraction of its parameters. The basic model represented in figure 2.8 is enough to study the dynamic behavior of a SPADs, a more complex and realistic model can be find in [72]. This basic model emulates the onset of

Table 2.1: Usual magnitudes for the SPAD circuit model of figure 2.8.

Parameter	Value	Unit
V_{BD}	10–40	V
R_{BD}	100–1000	Ω
C_{KA}	100–1000	fF
C_{AB}	1–10	fF
C_{KB}	1–2	pF

the avalanche with closing of the switch labeled as "Event". At that moment, the current starts flowing from cathode (K) to anode (A), discharging C_{KA}, until the voltage across the diode terminals reaches the breakdown voltage, V_{BD}. With the avalanche quenched, the circuit starts the recharge until the bias voltage is reached.

As it is evident, the circuit parameter values are very important, since they determine, along with the QRC, the time constant of the avalanche quenching and of the diode recharge, and the current magnitude of the avalanche. In the best case scenario, the designers have experimental data from which they can obtain all the required values. However, sometimes this ideal case may not be true. Without experimental measurements, designers can make use of the library cells provided by the foundry. In the worst case scenario, designers must use the reported values by other groups in order to have a SPAD model whose parameter values are in the proper order of magnitude. The common values of these parameters in CMOS technologies are summarized in table 2.1 [65, 73]. From now on, the following values are used: $V_{BD} = 18\,\text{V}$, $R_{BD} = 300\,\Omega$, $C_{KA} = 200\,\text{fF}$, $C_{AB} = 10\,\text{fF}$ and $C_{KB} = 1\,\text{pF}$. Some of these values have been extracted from the models provided by the foundry, while others have been obtained from the bibliography.

2.4.1 Passive Quenching and Recharge Circuit

The design of a PQRC consist in determining the size of the quenching resistor. Firstly, the value of the resistance is chosen to guarantee that the steady-state current flowing through the SPAD is below the self-quenching value, I_{quench}. Below this quenching current, the avalanche is not longer self-sustained and, eventually, it will be quenched, see figure 2.9. Although the value of this current depends on the technology and the junction, reported values area around 10–100 µA [65, 68, 53]. Even though the steady-state current through the SPAD is below the self-quenching value, the current peak associated to the discharge of C_{KA} through R_{BD} could be order of magnitude higher. As it is shown in figure 2.9-(a), if $R_Q \gg R_{BD}$ and the switch is closed, the current flowing through R_Q in steady-state is given by (2.7). This approach is equivalent to find the operating point of any circuit. Hence, R_Q should has a value around 60–600 kΩ

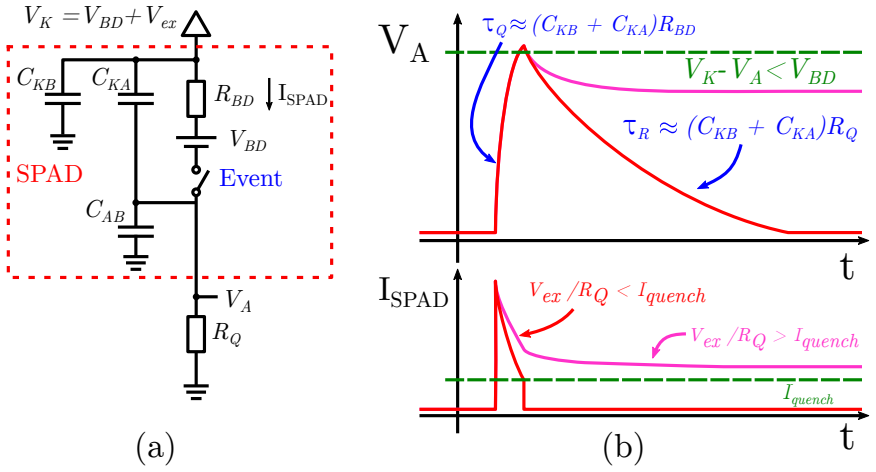


Figure 2.9: SPAD with passive quenching and recharge circuit (PQRC): (a) circuit schematic and (b) dynamic response, not into scale.

for $V_{ex} = 6$V.

$$I_{SPAD} \approx \frac{V_{ex}}{R_Q} \tag{2.7}$$

Secondly, it is important to estimate the quenching and recharge times of the SPAD. Analyzing the circuit in figure 2.9-(a), when the switch is closed and $R_Q \gg R_{BD}$, the quenching time constant is [65]:

$$\tau_Q = (C_{KA} + C_{AB}) \frac{R_Q R_{BD}}{R_Q + R_{BD}} \approx (C_{KA} + C_{AB}) R_{BD} \tag{2.8}$$

Finally, when the switch is open again, the recharge time constant can be calculated as:

$$\tau_R = (C_{KA} + C_{AB}) R_Q \tag{2.9}$$

From (2.8) and (2.9) it can be observed that the recharge time is much longer than the quenching time. Furthermore, the bandwidth of the circuit, i.e. maximum number of avalanches per unit of time that can be triggered without saturating the SPAD output, is function of R_Q, as expression (2.10) shows. At first glance, this could be compensated reducing V_{ex}, however, this has a negative impact on the PDP and jitter.

$$BW \propto \frac{1}{\tau_R + \tau_Q} \approx \frac{1}{(C_{KA} + C_{KB}) R_Q} = \frac{I_{SPAD}}{(C_{KA} + C_{KB}) V_{ex}} \tag{2.10}$$

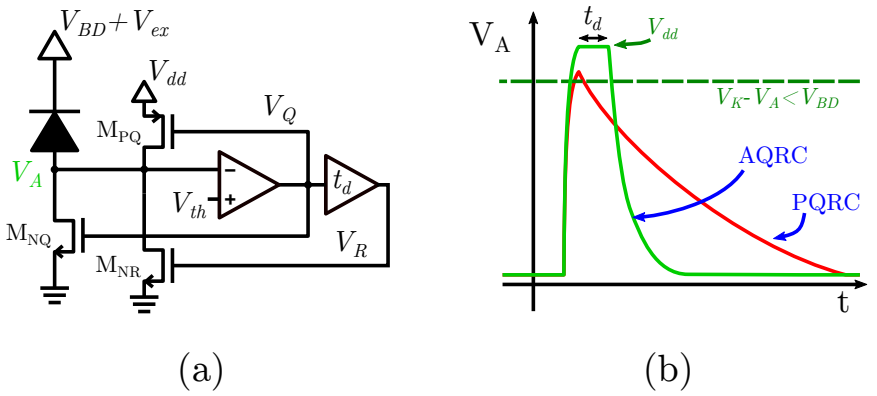


Figure 2.10: SPAD with active quenching and recharge circuit (AQRC): (a) circuit schematic and (b) dynamic response.

2.4.2 Active Quenching and Recharge Circuit

The working principle of an active quenching and recharge circuit is relatively simple as figure 2.10 illustrates. When the avalanche is triggered, the anode voltage (V_A) rises above the threshold voltage (V_{th}). Then, transistor M_{NQ} is turned off and M_{PQ} is turned on, pulling-up the anode voltage to V_{dd} and, hence, forcing a voltage across SPAD terminals smaller than the breakdown. Next, after delay (t_d), the transistor N_R is turned on, pulling the anode voltage down to ground. This resets the device to its initial operation point, waiting for another avalanche. As figure 2.10-(b) represents, the principal benefit of AQRC over PQRC is the control of the recharge time. This allows to increase the bandwidth without having a negative impact on the quenching of the avalanche. Furthermore, since the device spends more time out of Geiger mode and the deep level traps have more time to be discharged, the probability of after-pulsing is reduced.

The comparator and the delay element can be implemented in different ways depending on the specifications, from a few inverters as in [74, 75] to an operational amplifier [65, 76]. Nowadays, research in this field is very active and mainly focused on the reduction of the footprint to increase fill factor. For this reason, the following analysis will be qualitative and oriented to the design of essential transistors, M_{NQ}, M_{PQ} and M_{NR}.

Different design approaches, sometimes diametrically opposed, can be found in the bibliography. For example, in the sizing of the quenching transistor. For example, in [67], transistor M_{NQ} is sized to provide a high output resistance of 1.2 MΩ. On the contrary, other authors employ much more conductive transistors to better detect the avalanche current peaks [77]. This approach implies that the avalanche can only be quenched after the output of the comparator goes to zero and transistor M_{NQ} is switched off. Using this method, considering

I_{SPAD} = 0.5–1 mA [77] and $V_{ex} = V_{dd}$ = 3.3V, the size of M_{NQ} can be estimated with a quadratic model:

$$\frac{V_{ex}}{I_{SPAD}} = \frac{1}{R_Q} = \frac{1}{\frac{W}{L}\frac{\mu_n C_{ox}}{2}(V_{dd} - V_{th_n})^2\lambda} \Rightarrow$$
$$W = \frac{2I_{SPAD}L}{\mu_n C_{ox} V_{ex}(V_{dd} - V_{th_n})^2\lambda} \tag{2.11}$$

where R_Q is the output resistance of N_{NQ} transistor, $\mu_n = 0.03695\,\mathrm{m^2\,s/V}$, V_{th_n} = 0.5V, $\lambda = 0.02\,\mathrm{V^{-1}}$ and $C_{ox} = 10\times10^{-3}\,\mathrm{F/m^2}$. If the minimum length is used (L_{min} = 0.36µm), the transistor width is W = 1.5–3 µm. On the contrary, the approach presented in [67, 70] would yield a completely different result, that is, a transistor longer than wider. Considering (2.11), I_{SPAD} = 10–200 µA and W_{min} = 0.6µm, the transistor length is L = 0.6–5.7 µm. As it will be seen in chapter 5, the selection of the first method can explain the excessive power consumption and system instability.

Another point where special care must be taken is during the design of M_{NR} and M_{PQ}, to avoid instabilities related to a positive feedback, like oscillations or V_A stuck at V_{dd}. If the gate of M_{PQ} is simply driven by the comparator, M_{NR} must be more conductive than M_{PQ}, to pull-down the anode voltage to ground. To avoid any short circuit current between V_{dd} and ground when both transistors are active, the straightforward solution is to add a second PMOS transistor in series with M_{PQ} and connect its gate to V_R. However, this last solution may not be attractive for arrayed devices, because it reduces the fill factor by increasing the occupation area.

2.4.3 Passive Quenching Active Recharge Circuit

In the bibliography, some alternatives halfway between AQRC and PQRC are proposed to provide the simplicity of a passive quenching and the bandwidth of an active recharge [53]. Moreover, while the excess voltage is limited to V_{dd} in an AQRC architecture, the excess voltage in a passive quenching and active recharge circuit (PQARC) is limited by the transistor maximum operation voltages. For example, the PQARC architecture shown in figure 2.11-(a) is able to work with an excess voltage above V_{dd}. Thanks to the cascode transistor M_{NC} that reduces the voltage across the quenching transistor to less than the maximum V_{GS} and V_{DS} allowed by the technology [78]. However, the maximum excess voltage is limited by the breakdown of the drain-bulk junction of M_{NC}.

The operation principle of this topology is shown in figure 2.11-(b) and it can be divided into tree main parts: quenching, dead time and recharge. Once the avalanche is triggered, the current rises until it reaches the value set by the current mirror and the anode voltage reaches a value close to the excess voltage. This quenches the avalanche. Next, the anode voltage starts discharging

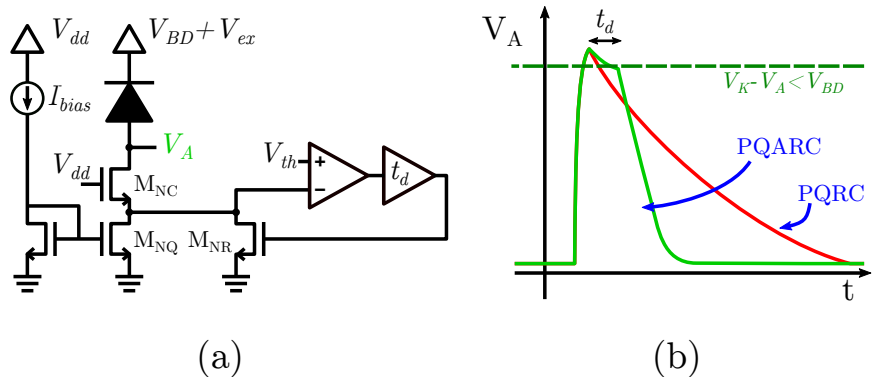


Figure 2.11: SPAD with cascode passive quenching active recharge circuit (PQARC): (a) circuit schematic and (b) dynamic response.

through the cascode and quenching transistor. The current is limited by the current mirror during a time t_d. After that, the reset transistor is switched on and the anode voltage is completely discharged. This topology is also known as double-slope passive quenching and active recharge circuit [53].

In terms of bandwidth and after-pulsing the dead time can be controlled as easy as in an AQRC, but since the anode voltage is not forced to supply, the power consumption is slightly smaller. All these advantages make this architecture very attractive for applications were large SPAD arrays are needed.

2.5 Proposed Active Quenching and Recharge Circuit

The schematic of the proposed circuit is illustrated in figure 2.12-(a) and it is similar to the presented in [79, 80]. In this case, each micro-cell is composed of several SPADs and AQRCs that share some common logic, monostable and reset circuit. Thus, the shared logic reduces the number of transistor per each AQRC and increments the fill factor at the expense of sharing the dead time among all the devices. The concept of sharing some logic among AQRCs was presented in [81] and the their effects on the SiPM performance will be analyzed in chapter 4.

The working principle of this structure is simple. If the enable signal E is V_{dd}, the circuit works as follows. As figure 2.12-(b) illustrates, when the avalanche is triggered, the rising edge of the anode voltage switches on the transistor M_{ND} and pulls-down the voltages V_{nQ} and V_{out}. Then, transistor M_{NQ} is switched off and M_{PQ} is switched on. This ties up the anode voltage to V_{dd} and quenches the avalanche. After the dead time t_d the reset signal rises and the anode voltage is pull-down. Then, the voltage V_{out} is set to V_{dd} by a pull-up. On the contrary,

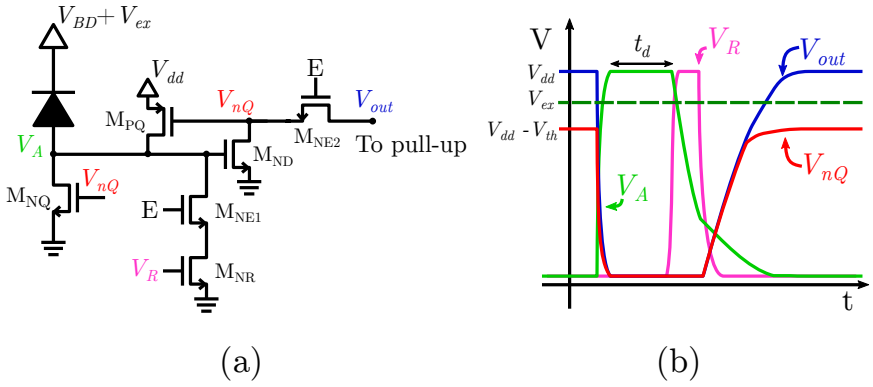


Figure 2.12: SPAD with proposed active quenching active recharge circuit: (a) circuit schematic and (b) dynamic response. The enable signal E is loaded into a SRAM during the configuration of the cell. The reset voltage V_R is generated with a monostable. The output voltage V_{out} is connected to a pull-up OR to sum the output of several SPADs.

Table 2.2: Transistor sizes for circuit of figure 2.12.

Transistor	W/L (µm/µm)	Transistor	W/L (µm/µm)
M_{NQ}	3/0.36	M_{ND}	14.4/0.36
M_{PQ}	2/0.36	M_{NE1}	4.6/0.36
M_{NR}	4.6/0.36	M_{NE2}	3.6/0.36

if the enable signal E is grounded, the anode voltage is stucked at V_{dd} after the first avalanche, because the reset path is open. The enable signal E can be loaded into a static random access memory (SRAM) close to the AQRC to turn off the noisy SPADs.

The transistor sizes are summarized in table 2.2. The sizing of M_{NQ} was done following the approach shown in section 2.4.2 [77]. This means that the maximum current flowing through the SPAD is 0.5–1 mA. M_{PQ} is sized to tie up the anode to V_{dd} in 0.2–0.5 ns and M_{NR} in series with M_{NE1} must be more conductive than M_{PQ} to discharge the anode to ground. Finally, M_{ND} is wide to make the front-end sensitive for low excess voltages. This is equivalent to set a low threshold voltage at the comparator.

Figure 2.13 represents, as an example, the schematic and layout of a micro-cell with 4 SPADs. The design is compact and allows to be abutted with more micro-cell thanks to the shared deep N-well. The output of several micro-cells can be OR connected to a counter, this way the number of detected photons can be digitally counted. Despite all the implemented logic in the micro-cell, its fill factor is 30 %.

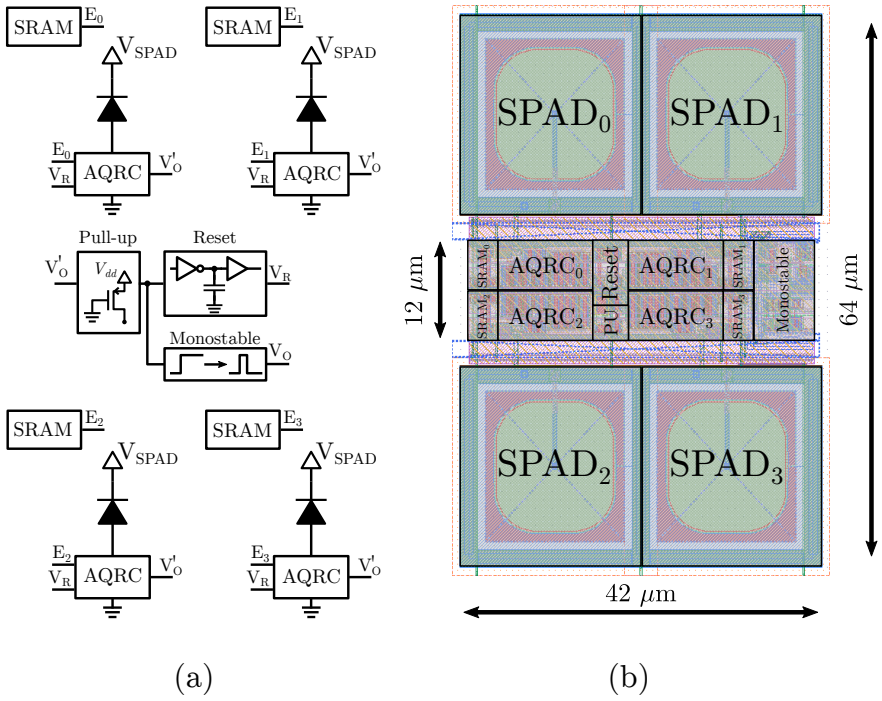


Figure 2.13: Micro-cell with 4 SPADs sharing logic: (a) circuit schematic and (b) layout.

2.6 Conclusion

In terms of SPAD structures, it has been shown that the perfect structure does not exist, because the final choice depends on the application and several trade-offs. As it was explained, a superficial structure has higher sensitivity to blue, more DCR, less jitter and cross-talk, while a deep junction has higher sensitivity to red, less DCR, more jitter and cross-talk. To test the performance of technology, a superficial structure composed of p+ diffusion and N-well layers (PP/NW3V3) and a deep junctions composed of P-well and Deep N-well layers (PW3V3/DNW).

Nowadays, the research in quenching and recharge circuits is still very active and mainly is pushed by the specifications of new applications. Here, the most common architectures were described to provide an insight in the main trade-offs and highlight the most critical design points. The working principle of proposed AQRC is described.

CHAPTER 3

TIME-TO-DIGITAL CONVERTER DESIGN

In order to accurately timestamp the impact of gamma photons in a scintillator crystal, an in-pixel TDC is needed. One of the most important metrics for TDCs is the time resolution which, in principle, is determined by technology. The minimum detectable delay is that of an inverter. However, finer resolutions can be achieved with techniques like, for instance, pulse stretching, Vernier delay lines, time amplification or multi-path gated ring oscillator. However, these architectures are unsuitable for those applications in which specifications demand the incorporation of in-pixel TDC, because of their large area, conversion time and power consumption. For instance, if a high frame rate is required, events need to be timestamped in parallel. For that, a TDC with not only a high resolution but also low area occupation and power consumption is needed. TDCs based on VCRO offer a good trade-off between time resolution, conversion rate, area and power. They are essential in a large range of applications, like all-digital phase locked loop (ADPLL)s for communications; ToF estimation in telemetry and particle identification; and bio-medicine, including FLIM and PET.

The main goal of this chapter is to justify the election of VCRO-based TDC and study analytically the main FOMs for this architecture. This chapter is organized as follows. The working principle of TDCs based on VCRO are presented in section 3.2. Secondly, a VCRO is studied as a core of the proposed TDC topology and expressions to estimate the oscillation frequency, the phase noise and other relevant FOMs are obtained. This study is carried-out for the two delay stages most widely employed in literature. Finally, in section 3.4, the implemented converter is described, putting special emphasis in the main block of the converter, like, the ripple counter, the transmission gates and the

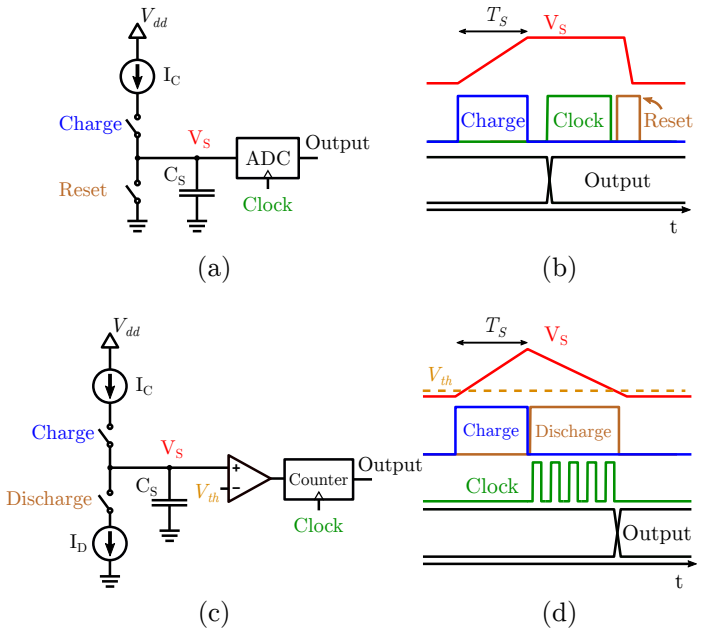


Figure 3.1: (a) Simplified block diagram of single-slope analog TDC, (b) signal diagram of single-slope analog TDC, (c) simplified block diagram of double-slope analog TDC and (d) signal diagram of double-slope analog TDC.

decoder.

3.1 TDC Architectures

Generally, the TDC architectures can be divided into to main groups, analog and digital. The working principle of analog single-slope TDCs is based on the integration of a constant current, I_C, into a capacitor, C_S, during the sampling time interval T_S. figure 3.1-(a,b) represents the simplified schematic and the signal diagram of a single-slope TDC. The voltage at the terminals of the sampling capacitor is proportional to the sampling interval as (3.1) shows:

$$I_C = C_S \frac{dV_S}{dt} \Rightarrow \int_0^{T_S} I_C dt = C_S \int_0^{T_S} dV_S \Rightarrow T_S = \frac{V_S\, C_S}{I_C} \tag{3.1}$$

This allows an analog to digital converter (ADC) to perform the conversion and provide the equivalent digital word for T_S. As the reader may expect, this architecture suffers from mismatch and PVT variations in deep sub-micron technologies. An obvious approach to circumvent these limitations is to apply a principle similar to that used in double-slope ADC converters. Thus, as figure 3.1-(c,d) illustrates, a constant current, I_C, is integrated into the capacitor,

C_S, during the sampling time, T_S. Then, the capacitor is discharged with a constant current, I_D, during a time, T_D. This discharge has a smooth slope than the charge slope. As a result, the sampling time is proportional to the discharge time and the ratio of currents:

$$T_S = T_D \frac{I_D}{I_C} \tag{3.2}$$

the discharge time can be digitized with a comparator and a counter of clock cycles as figure 3.1-(d) represents. Although this architectures presents a good time resolution and measuring range, the area occupied by the capacitor and the conversion time make this alternative no attractive for in-pixel implementations.

As for digital TDC architectures, their working principle is based on the comparison between the sampling time and the propagation delay of several logic gates. Hence, the output of the conversion is proportional to the number of logic gates through which the signal propagates. This, in principle, means that the minimum achievable temporal resolution by these converters is limited by the gate delay. As figure 3.2-(a) illustrates, this is the case of delay line architectures, in which the delay between the signal propagated through the buffers, T_{start}, and the edge of T_{stop} is recorded at the flip-flops with thermometric code. Here, the race condition between two signals is a necessary feature in this kind of circuits, despite being a problem like in most circuits. In order to increase the measuring range, the easiest solution is to implement a ring oscillator, as figure 3.2-(b) illustrates. Thus, the coarse conversion is given by the number of oscillations done by the ring during T_S, while the fine conversion is a function of the internal phases at which the ring is stopped. As it will be discussed later, this architecture shows a good trade-offs between area, power, conversion range, resolution and conversion time.

Different architectures can be found in the literature for applications that require temporal resolution finer that the delay og a single logic gate. Nevertheless, resolution improvement comes at the expense of an increment in power consumption, area occupation and, usually, the conversion time. As a consequence, they are not very attractive for in-pixel integration. The principal sub gate-delay TDC architectures are described in the following.

Figure 3.3 shows the block diagram of a Vernier delay line. Compared to the conventional delay line, this architecture have a second delay line in the propagation path of the stop signal, T_{stop}. Since the two delay lines have different delay stages, for instance, slow elements, τ_s, in the start path and fast elements, τ_f, in the stop path, the converter resolution is $\tau_s - \tau_f$. The conversion provides a thermometric code in which the last active flip-flop indicates the point where the start signal is caught by the stop signal.

The pulse shrinking architecture represented in figure 3.4, employs two inverters with different rising edge delay, t_{dr}, and falling edge delay, t_{df}, to reduce the width of the pulse T_s. This means that every time the pulse is prop-

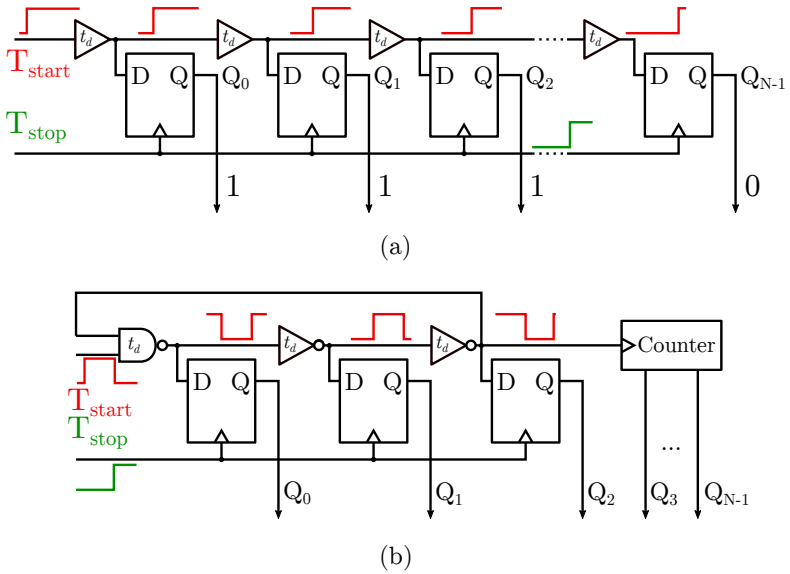


Figure 3.2: Simplified block diagram of digital TDC architectures: (a) delay line, (b) VCRO.

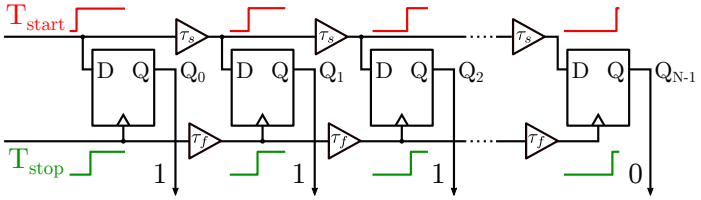


Figure 3.3: Simplified block diagram of Vernier delay line TDC.

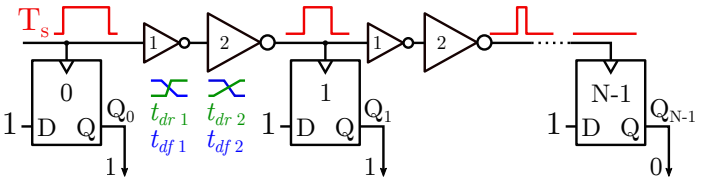


Figure 3.4: Simplified block diagram of pulse shrinking TDC.

agated through a delay stage its width is reduced a value equal to the time resolution [82]:

$$T_{shr\ res} = (t_{dr\ 2} - t_{df\ 2}) - (t_{dr\ 1} - t_{df\ 1}) \tag{3.3}$$

where $T_{shr\ res}$ is the resolution time, $t_{dr\ 2}$ is the rising edge delay of stage 2, $t_{df\ 2}$

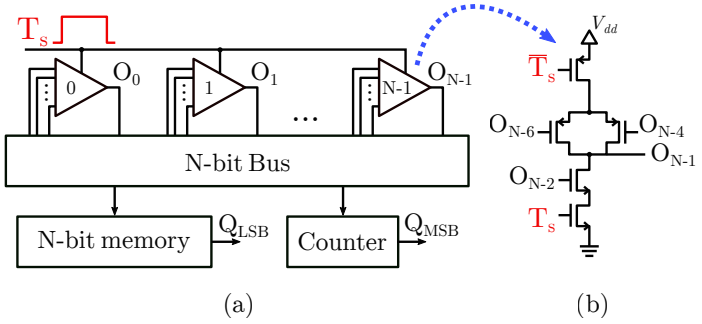


Figure 3.5: (a) Simplified block diagram of multi-path TDC and (b) schematic of delay stage.

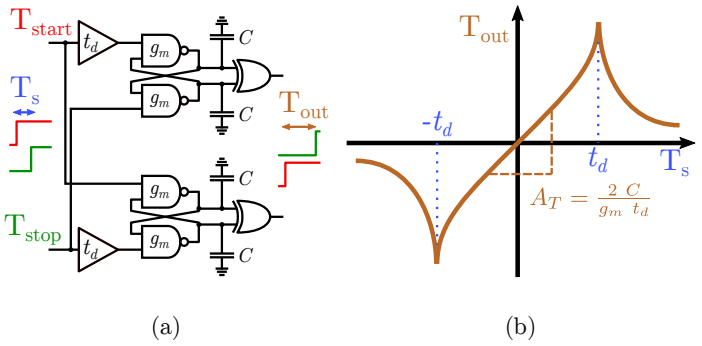


Figure 3.6: (a) Simplified block diagram of time amplification TDC and (b) transfer characteristic of time amplification.

is the falling edge delay of stage 2, $t_{dr\,1}$ is the rising edge delay of stage 1 and $t_{df\,1}$ is the falling edge delay of stage 1. The conversion finishes when the pulse vanishes in the delay path, because the signal is below the noise margin of the inverters. This makes the architecture sensitive to PVT variations [82].

Multi-path gated ring oscillator architecture is shown in figure 3.5-(a). Multi-path architectures are implemented with delay stages with multiple inputs, each one connected to different delayed phases, see figure 3.5-(b). With this, a reduction of the propagation delay per stage is achieved [83]. However, the design of such oscillators is not trivial, since several oscillation frequencies, phase shifts and phase noises are possible because of the existence of different oscillation modes [84]. A careful design must be carried out to avoid unwanted oscillation modes. As the ring oscillator architecture shown before, this structure also uses the oscillator phases and a counter to obtain the fine and the coarse conversion, respectively.

In time amplification architectures, the width of narrow pulses is increased thanks to the metastability of positive feedback comparators, see figure 3.6-(a).

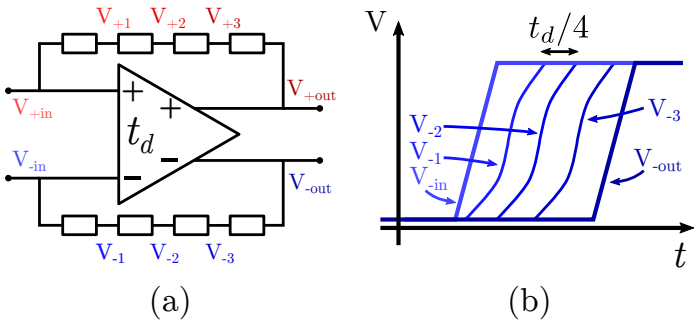


Figure 3.7: (a) Delay stage of local passive interpolation TDC and (b) interpolated signals between input and output edges.

The small-signal time amplification is given by [85]:

$$A_T = \frac{2\,C}{g_m\,t_d} \tag{3.4}$$

where C is the load capacitance of NAND gates, g_m is the small-signal transconductance of NAND gates during metastability and t_d is the delay time that extends the metastability input range. Because of the non-monotonic response of the transfer characteristic, see figure 3.6-(b), the input range must be between $\pm t_d$. The main drawback of this technique is its limited measuring range and linearity. Consequently, this architecture is commonly used as a part of pipeline or fine-coarse converters [85, 86].

In addition to all these architectures, local passive interpolation architectures are also part of the sub gate-delay family, without having some of their drawbacks like long conversion times or unwanted oscillation modes [87]. Local passive interpolation architectures are equivalent to delay lines, but each delay stage has a voltage divider between input and output, as figure 3.7-(a) illustrates. This voltage divider creates a copy of the input signal in their internal nodes as figure 3.7-(b) shows. As a result, the voltage interpolation increases the number of signal phases and, hence, the resolution. One of the main drawbacks of this architecture is the power consumption because of the cross current at the delay stage during interpolation [87].

3.2 Proposed Architecture

As it was discussed in the previous section, for a given technology, the selection of the architecture determines the compromise between time resolution, area and power consumption. The TDCs based on VCROs are generally chosen for in-pixel implementation because of several reasons. First of all, as the oscillation frequency is a function of the delay of individual stages of the oscillator,

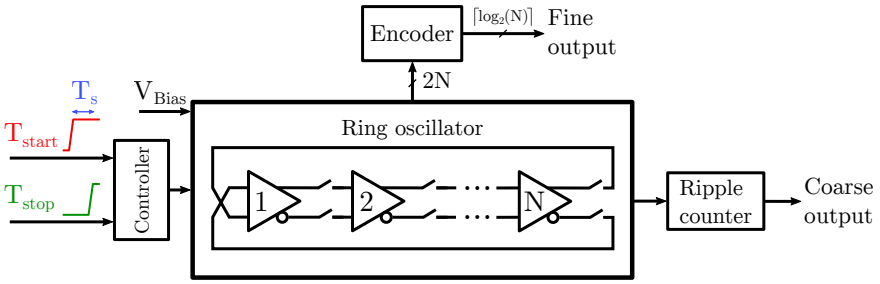


Figure 3.8: Simplified block diagram of a voltage controlled ring oscillator (VCRO) time to digital converter (TDC).

the time resolution can be controlled by controlling the stage delay, T_d. That can be done through, for instance current-starved stages or voltage-controlled transmission gates. The main advantage of voltage-controlled transmission gates is the possibility to adjust the oscillation frequency linearly to a reference value through a control loop, which reduces the effects of PVT variations [88]. On second place, the small area footprint of this architecture makes it possible to integrate the TDC within a pitch size as small as $50\,\mu m \times 50\,\mu m$ [89]. Finally, its conversion time is small enough to allow sampling frequencies in the order of tens of megahertz [90]. These are the main reasons why the VCRO architecture is chosen.

Figure 3.8 displays a simplified block diagram of the proposed VCRO TDC. The time interval to be sampled, T_s, is determined by the difference between the rising edges of the start and stop signals, T_{start} and T_{stop}, respectively. The controller block prevents that any spurious pulse re-triggers the oscillator. The voltage V_{Bias} modifies the resolution of the converter by controlling the stage delay. Once the VCRO is stopped, the ripple counter retains the number of oscillations done by the VCRO, constituting the coarse conversion of the sampled time interval. Simultaneously with the rising edge of the stop signal, the internal phases of the VCRO are encoded into additional bits, that provide the fine conversion. Notice that, unlike a single-ended ring oscillator where there must be an odd number of stages, in differential and pseudo-differential ring oscillators, the number of stages can be even. In order to fulfill Barkhausen criteria, however, a wire inversion has to be placed [91].

3.3 Study of Delay Stages

Nowadays, VCROs can be found in a wide range of applications, like ADPLLs for communications [90]; direct ToF estimation for telemetry [92, 93] and high energy physics (HEP) [94]; and bio-medicine, including FLIM [95] and PET [44]. In particular, flash 3D imaging based on direct ToF estimation relies on in-pixel TDCs. They need to be compact and power efficient to fit into a small

pixel pitch. TDC architectures employing pseudo-differential VCROs as time interpolators fulfill these requirements. At this point, the design of the VCRO becomes central. In the following sections, an effective analysis is proposed to optimize the design parameters of the pseudo-differential VCRO to maximize the overall performance of in-pixel TDCs.

The main building block of VCROs is the delay cell which can be single ended, differential or pseudo-differential. Unlike differential delay cells, single ended and pseudo-differential delay cells achieve rail-to-rail outputs, because they lack the biasing current source. This rail-to-rail output swing features lower jitter due to smaller thermal noise associated with larger slew rates [91, 96]. Besides, they do not require any static power which makes them more power efficient. Differential stages with non-saturated outputs provide excellent common mode rejection [91]. However, they need a bias current to keep the operating point in a linear region, this implies higher power consumption. Pseudo-differential delay cells have worse rejection to supply and ground noise than differential rings, but better than single ended delay cells. Finally, differential and pseudo-differential rings can be built with even or odd number of stages, even rings are useful in clock recovery or phase locked loops when quadrature outputs are needed [97]. For these reasons, most of the implementations in the literature are based on differential [89] and pseudo-differential rings [98, 99].

Since differential rings with non-saturated outputs work in the linear region, it is almost straightforward to make an estimation of the oscillation frequency based on the transfer function of the oscillator using small-signal analysis. Instead, pseudo-differential oscillators are more difficult to predict because of their strongly non-linear behavior. Despite these non-linearities, it is possible to make a first order approximation and calculate the pole frequency of the stage by a small-signal approach [100]. This approximation only remains valid while the non-linearities of the output signal are small. This can be achieved choosing the proper number of delay stages, for instance, in a ring with a large number of delay stages, the output is hard saturated (or hard clipped), because it spends a large fraction of the period in the supply rails. However, in a ring with small number of delay stages, the output is soft saturated, since it spends a smaller fraction of the period in the supply rails. In order to avoid the non-linearities and high harmonics inherent to hard saturation, simulations are done with a 3-stage ring oscillator. Otherwise, one should work with tricky expressions, not very suitable to highlight the trade-off between stages.

Figure 3.9-(a) shows the pseudo-differential stage with cross coupled PMOS (XCP) transistors, while the pseudo-differential stage with cross-coupled inverters (XCI) is shown in figure 3.9-(b). These two pseudo-differential stages are widely used in the literature: XCP stages can be found in [89, 101, 102, 103]; while XCI stages can be found in [98, 99, 104, 105]. The main difference between them is that C_L is charged by cross-coupled element in the XCP stage, while C_L is charged by delay elements in the XCI stage. In other words, in the XCP stage, PMOS transistors have two roles: charging C_L and providing positive-feedback.

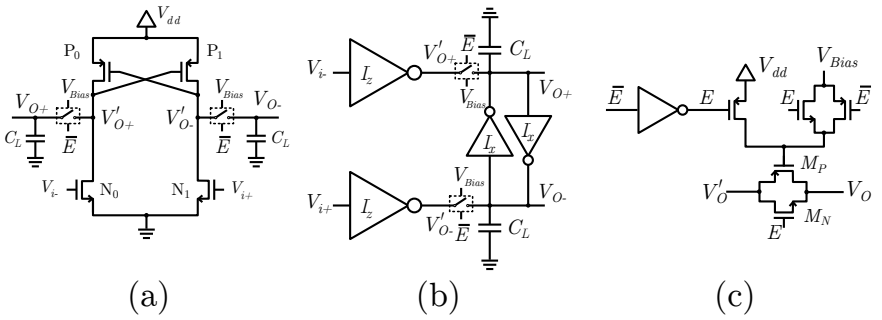


Figure 3.9: (a) Pseudo-differential stage with cross-coupled PMOS (XCP), (b) pseudo-differential stage with cross-coupled inverters (XCI) and (c) tunable resistor. C_L is the capacitive load, V_{Bias} is the tuning voltage, V_o' and V_o are the terminals of the tunable resistor, E and $\overline{E}$ is the enable and its complementary, respectively.

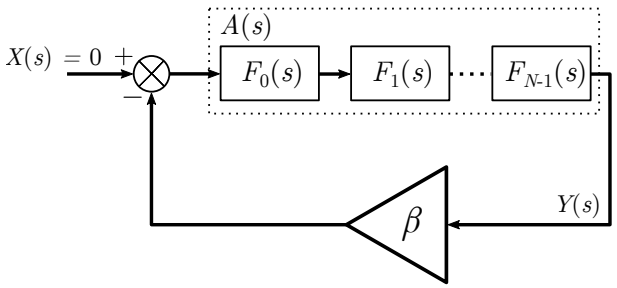


Figure 3.10: Simplified block diagram of N-stages ring oscillator.

While in XCI, these tasks are done by delay inverters and cross-coupled inverters, respectively.

The frequency tuning of these rings is achieved by means of a voltage controlled resistor between each delay stage [99]. The schematic of the tunable resistor is shown in figure 3.9-(c). The tuning voltage, V_{Bias}, changes the conductance of M_P transistor, while the voltage gate of M_N is kept constant. This means that, for a value of $V_{Bias} = V_{dd}$, M_P is in cut-off, while M_N is still active, and oscillations are possible. The main advantage of this method is to control the oscillation frequency almost linearly with the tuning voltage as will be shown in the following sections.

3.3.1 Oscillation Frequency

Single-ended, differential with clipped outputs and pseudo-differential ring topologies are highly non-linear systems, due to the fact that the transistor do not operate around the same operation point, e.g., saturation, linear or cut-off, but rather, it fluctuates from one region to the other during its operation. For

that reason, the study and analysis of these ring oscillators is not trivial. To circumvent this issue, the feedback loop can be divided into its linear time invariant and non-linear time invariant components. Doing so, the oscillation frequency can be estimated by means of frequency domain linear analysis, whereas the non-linear component determines the output swing [100]. This assumption is valid as long as the amount of time that the output remain hard saturated is negligible. In other words, the output signal is more sinusoidal than squared and it has less harmonics. Therefore, in this section a small-signal analysis is performed to estimate the oscillation frequency of the ring. In figure 3.10, the simplified block diagram of a ring oscillator with N stages is shown. The closed loop transfer function, $H(s)$, of the whole system has the following form:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{A(s)}{1 + A(s)\beta} \tag{3.5}$$

Barkhausen criteria is a necessary condition for steady state oscillation but not sufficient for starting-up oscillations. To fulfill Barkhausen criteria at the angular oscillation frequency, ω_0, the characteristic equation of (3.5) becomes:

$$1 + A(s)\beta = 0 \Rightarrow A(j\omega_0)\beta = -1 \tag{3.6}$$

where $A(j\omega_0)$ is the transfer function of N stages and β is the feedback transfer function, the time invariant non-linear component, that is considered equal to 1 during this step of the analysis. Therefore, the oscillation condition can be rewritten as:

$$|A(j\omega_0)|_{\angle\phi} = \left|F(j\omega_0)\right|^N_{\angle\pi} = 1_{\angle\pi} \tag{3.7}$$

for simplicity, the transfer function of each delay stage, $F(j\omega_0)$, has a single dominant pole at ω_p:

$$F(j\omega_0) = \frac{A_0}{1 + j\frac{\omega_0}{\omega_p}} \tag{3.8}$$

applying (3.8) into (3.7), the following equations are obtained:

$$\begin{gathered} \left[\frac{A_0}{\sqrt{1+\left(\frac{\omega_0}{\omega_p}\right)^2}}\right]^N = 1 \\ \phi = N\arctan\left(\frac{\omega_0}{\omega_p}\right) = \pi \end{gathered} \tag{3.9}$$

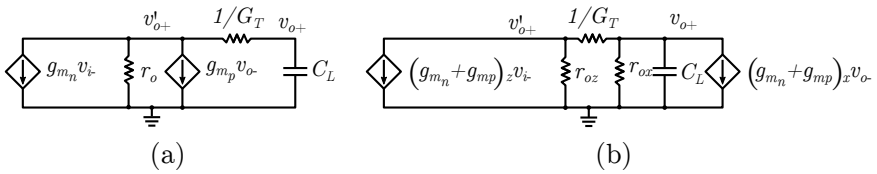


Figure 3.11: (a) Small-signal model of the half pseudo-differential XCP stage and (b) small-signal model of the half pseudo-differential XCI stage.

solving the expression (3.9) for the oscillation frequency, f_0:

$$f_0 = \frac{\omega_0}{2\pi} = \frac{\omega_p}{2\pi}\tan\left(\frac{\pi}{N}\right) \tag{3.10}$$

Expression (3.10) illustrates that the oscillation frequency in the small-signal model, f_0, is function of two parameters: the pole frequency, ω_p; and the number of stages, N. Consequently, it is fundamental to study the small-signal model of each stage to find which of them provides the highest oscillation frequency.

Pseudo-Differential Stage with Cross-Coupled PMOS

Figure 3.11-(a) represents the small-signal model of the XCP stage. The transfer function of this stage is:

$$F_{XCP}(s) = -\frac{g_{m_n}}{g_{m_p} - g_o - sC_L\left(1 + R_T g_o\right)} \tag{3.11a}$$

$$\omega_p = \frac{g_o - g_{m_p}}{C_L(1 + R_T g_o)} \tag{3.11b}$$

where $F_{XCP}(s)$ is the transfer function of the pseudo-differential stage with cross-coupled PMOS, g_o is the output conductance of the combination of M_{P0} and M_{N0} transistors, g_{m_p} is is the transconductance of M_{P0} which is in parallel to the output conductance g_o, R_T is the resistance of the tunable resistor described by (3.13), and the capacitor C_L can be approximated as:

$$C_L = (C_{gs} + C_{gb} + C_{gd})_n + (C_{gs} + C_{gb} + 2C_{gd})_p + (C_{gs} + C_{sb})_{M_N} + (C_{ds} + C_{db})_{M_P} \tag{3.12}$$

where the subindex M_N and M_P stand for the NMOS and PMOS of the tuning resistor, see figure 3.9-(c). G_T can be approximated as the conductances of the tunable resistor in linear region:

$$G_T = \frac{1}{R_T} = \beta_{M_N}\left(V_{dd} - V_o - V_{th_n}\right) + \beta_{M_P}\left(V'_o - V_{Bias} - |V_{th_p}|\right) \tag{3.13}$$

where $\beta = \mu C_{ox}\frac{W}{L}$, V_{th} is the threshold voltage, V_o is the output voltage of the delay stage and input of the tunable resistor, V_o' is the voltage at the other terminal of the tunable resistor, and V_{Bias} is the tuning voltage. Expression (3.13) is valid while $V_o' < V_{dd} - V_{th_n}$ and $V_o > V_{Bias} + |V_{th_p}|$ [99]. Also, the output conductance of this stage, g_o, in saturation is:

$$g_o = g_{ds_n} + g_{ds_p} \tag{3.14}$$

where g_{ds_n} and g_{ds_p} are the small-signal common-source output conductances of M_{P0} and M_{N0} transistor, respectively. Finally, substituting (3.14) and ω_p from (3.11b) into (3.10), the oscillation frequency can be defined as:

$$f_0 = \frac{G_T(g_{ds_n} + g_{ds_p} - g_{m_p})}{2\pi\left(G_T + g_{ds_n} + g_{ds_p}\right)C_L}\tan\left(\frac{\pi}{N}\right) \tag{3.15}$$

As it was shown in the previous section, the main function of PMOS transistors is to charge C_L up to V_{dd} through the conductance g_{ds_p}. Furthermore, as expression (3.15) shows, an increase of the PMOS strength not only increases g_{ds_p} but also g_{m_p}, this could lead to a reduction of f_0. Moreover, if it is considered that $G_T \ll g_{ds_n} + g_{ds_p}$, the oscillation frequency can be simplified to a linear function of V_{Bias}, if expression (3.13) remains valid.

Pseudo-Differential Stage with Cross-Coupled Inverters

figure 3.11-(b) represents the small-signal model of pseudo-differential stage with cross-coupled inverters. The transfer function of this stage is:

$$F_{XCI}(s) = -\frac{(g_{m_n} + g_{m_p})_z}{(g_{m_x} - g_{ox} - sC_L)\left(1 + R_T g_{oz}\right) - g_{oz}} \tag{3.16a}$$

$$\omega_p = \frac{G_T g_{o_z} + (G_T + g_{o_z})(g_{o_x} - g_{m_x})}{(G_T + g_{o_z})C_L} \tag{3.16b}$$

where the subindex z and x stand for the delay inverters, I_z, and the cross-coupled inverters, I_x, respectively. $F_{XCI}(s)$ is the transfer function of the pseudo-differential stage with cross-coupled inverters, g_{o_z} is the output conductance of the delay inverters, g_{o_x} is the output conductance of the cross-coupled inverters, $g_{m_x} = (g_{m_n} + g_{m_p})_x$ is transconductance of the cross-coupled pair, R_T is the resistance of the tunable resistor described by (3.13), and the capacitor C_L in saturation can be approximated as:

$$\begin{aligned} C_L = {} & (C_{gs} + C_{gb} + C_{gd})_{nz} + (C_{gs} + C_{gb} + C_{gd})_{pz} + (C_{gs} + C_{gb} + 2C_{gd})_{nx} \\ & + (C_{gs} + C_{gb} + 2C_{gd})_{px} + (C_{gs} + C_{sb})_{M_N} + (C_{ds} + C_{db})_{M_P} \end{aligned} \tag{3.17}$$

And the output resistances of this stage in saturation are,

$$\begin{aligned} r_{oz} &= \frac{1}{(g_{ds_n} + g_{ds_p})_z} \\ r_{ox} &= \frac{1}{(g_{ds_n} + g_{ds_p})_x} \end{aligned} \tag{3.18}$$

Finally, substituting (3.18) and ω_p from (3.16b) into (3.10) the oscillation frequency can be expressed as:

$$f_0 = \frac{\frac{G_T(g_{ds_n}+g_{ds_p})_z}{G_T+(g_{ds_n}+g_{ds_p})_z} + (g_{ds_n} + g_{ds_p})_x - (g_{m_n} + g_{m_p})_x}{2\pi C_L} \tan\left(\frac{\pi}{N}\right) \tag{3.19}$$

From expression (3.19) it can be observed that the contribution of delay inverters is independent of the negative resistance of the cross-coupled pair. Thus, it is possible to increase the strength of delay inverters respect to cross-coupled inverters without increasing the positive feedback term. This guarantees faster charges and discharges of C_L. In comparison with (3.15), expression (3.19) shows that this stage have one degree of freedom more because of its larger number of transistors. As an example, if delay inverters are sized to make $(g_{ds_n} + g_{ds_p})_z$ of (3.19) equal to $g_{ds_n} + g_{ds_p}$ of (3.15) and C_L is imposed equal in both expression, then g_{m_p} in (3.15) is always higher than $(g_{m_p} + g_{m_n})_x$ in (3.19), because cross-coupled inverters can be implemented with minimum size transistors. As a result, this suggests that a ring implemented with this stage has higher oscillation frequency.

Simulation Results

In order to validate the prediction on the oscillation frequency, two VCROs with 3 stages have been designed and simulated in 110 nm CIS technology. To make a fair comparison, both rings have been sized to have similar C_L at nominal size and $V_{Bias} = 0$V (see table 3.1). The data used to verify the accuracy of the models is obtained from fitting DC operating point simulation results to simulated oscillation frequency, unless otherwise indicated. The delay stages were connected as unity gain amplifiers. The input and the output of the tunable resistor were connected between V_{dd} and ground, respectively, for the DC operating point calculation. Simulated oscillation frequency, power consumption and phase noise are obtained with periodic steady-state analysis.

figure 3.12 shows the oscillation frequency of a ring with 3 pseudo-differential XCP stages as a function of bias voltage, V_{Bias}. Solid line represents the simulated oscillation frequency, while dashed line represents the frequency applying (3.15). At $V_{Bias} = 0$V and nominal sizes, this stage have $g_{ds_n} + g_{ds_p} = 180\mu$S, $G_T = 160\mu$S and the transconductance of PMOS transistors is set as $g_{m_p} = 38\mu$S

Table 3.1: Transistor nominal sizes and load capacitor.

	PD XCP	PD XCI		Tunable
		I_z	I_x	Resistor
W_p/L_p (μm/μm)	2.6/0.11	3.6/0.11	0.49/0.11	2.6/0.11
W_n/L_n (μm/μm)	3.1/0.11	1.1/0.11	0.15/0.11	1.3/0.11
C_L @ $V_{Bias} = 0$V (fF)	14.5	14.5		-

to produce the best fit. Under these conditions, the oscillation frequency is 1.2 GHz. Despite the good accuracy of the models at low V_{Bias}, an underestimation of G_T leads to a discrepancy between models and simulations at high V_{Bias}. As figure 3.13 illustrates, G_T behaves linearly if $V_{Bias} < 0.8$V, as (3.13) states. However, for values of $V_{Bias} \geqslant 0.8$V, the conductance is almost 0µS, because the transistors N_T and P_T are in the cut-off region. This cannot happen during oscillations since the transistor N_T is still conductive. Also, as (3.12) points out, some of the small-signal capacitors that compose C_L are function of V_{Bias}. Moreover, as it is expected, the values obtained from DC operating point of the delay stage are independent of V_{Bias} since the tunable resistor is the only element connected to V_{Bias}.

Figure 3.14 illustrates the variation of the oscillation frequency as a function of the ratio between PMOS and NMOS transistor width, W_p/W_n, at minimum lengths as table 3.1 shows. The discrepancy between model, f_0, and simulations, f_{sim}, at high W_p/W_n is attributed to an increment of non-linearities due to the rise of positive feedback term. It can be observed that in the small-signal model, a maximum separates the frequencies that are dominated by high output resistance at small W_p/W_n, and the frequencies that are dominated by high capacitive component, C_L, at high W_p/W_n (see Figs. 3.14 and 3.15).

Moreover, figure 3.16 confirms that the small-signal model of the XCI stage is able to estimate the oscillation frequency when the strength of cross-coupled inverters is increased. Even when the size ratio between cross-coupled and delay inverters, W_c/W_d, is close to 1. At $V_{Bias} = 0$V and nominal sizes, $W_c/W_d = 0.14$, this stage have $(g_{ds_n} + g_{ds_p})_z = 280$µS, $(g_{ds_n} + g_{ds_p})_x = 35$µS, $G_T = 160$µS and the transconductance of cross-coupled inverters is set as $(g_{mp} + g_{mn})_x = 25$µS to produce the best fit. Under these conditions, the oscillation frequency is 2.1 GHz. Solid line represents the simulated oscillation frequency of the ring, while dashed line represents the frequency applying (3.19). As it was previously discussed, expression (3.19) predicts that rings implemented with this stage topology have higher oscillation frequency, since the positive feedback

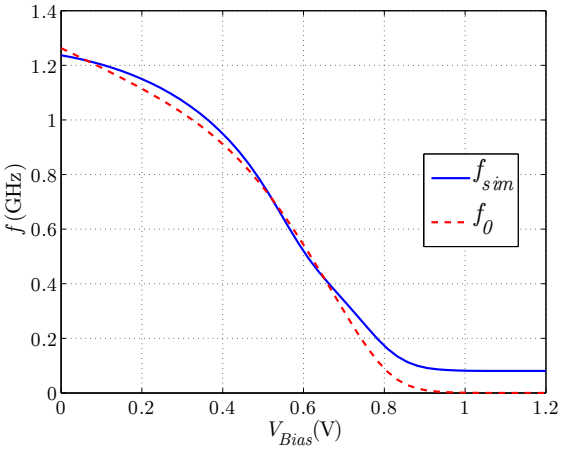


Figure 3.12: Oscillation frequency of a ring with 3 XCP stages as a function of V_{Bias}. Solid lines represents the simulated frequency of the ring, dashed line shows the results of small-signal model.

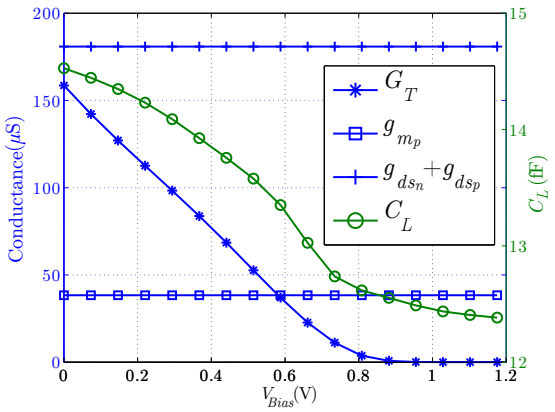


Figure 3.13: G_T, g_{m_p}, $g_{ds_n} + g_{ds_p}$ and C_L of the XCP stage as a function of V_{Bias}.

term, $(g_{m_p} + g_{m_n})_x$, is smaller than the conductances of delay inverters, $(g_{dsn} + g_{dsp})_z$.

3.3.2 Phase Noise

The noise performance of oscillators is another key parameter that must be taken into account during the design phase. In every oscillator, phase noise is responsible of the degradation of timing performance. The phase noise of a signal $V(t)$ arise from the random changes of the signal phase, ϕ [106, 107, 108, 109]:

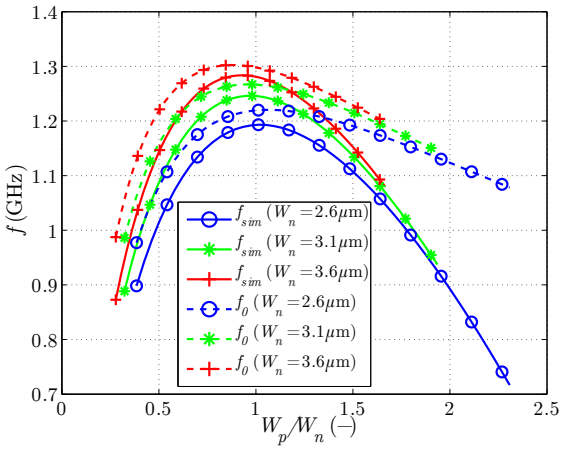


Figure 3.14: Oscillation frequency of a ring with 3 XCP stages as a function of W_p/W_n with $V_{Bias} = 0\,\text{V}$. Solid lines represent the simulated frequency of the ring, dashed lines show the results of small-signal model.

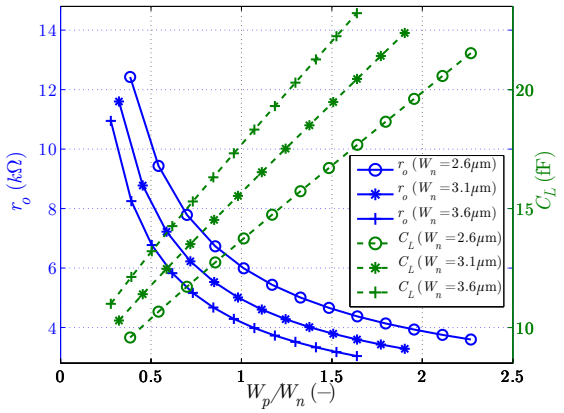


Figure 3.15: r_o and C_L of the XCP stage as a function of W_p/W_n with $V_{Bias} = 0\,\text{V}$.

$$V(t) = A\,\sin\left[2\pi\ f_0\ t + \phi(t)\right] \tag{3.20}$$

where f_0 is the signal frequency, $A(t)$ and $\phi(t)$ stand for the amplitude and phase fluctuations due to noise. As a result of these fluctuations, the spectrum of the signal is spread around the signal frequency. As for the oscillators considered during this analysis, the flicker $(1/f)$ noise of the tunable resistor is the principal contributor to phase noise [99, 109].

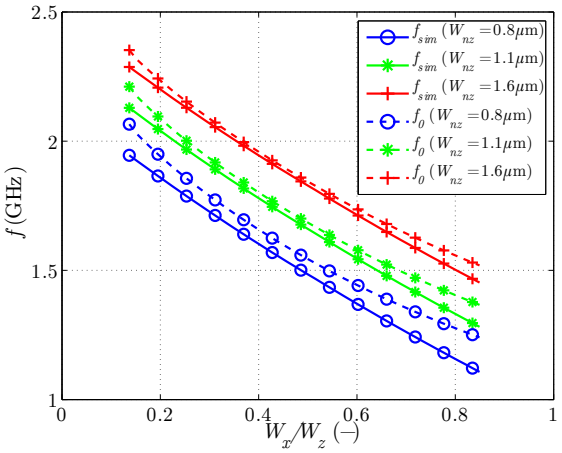


Figure 3.16: Oscillation frequency of a ring with 3 XCI stage as a function of W_x/W_z with $V_{Bias} = 0$V and $W_{pz} = 3.6$µS. Solid lines represent the simulated frequency of the ring, dashed lines show the results of small-signal model.

Pseudo-Differential Stage with Cross-Coupled PMOS

The single side band phase noise power spectral density caused by flicker noise in the ring oscillator can be defined as [99, 109],

$$\begin{aligned} \mathscr{L}(f) &= \frac{S_{f_0}^{1/f}}{4f^2} = \frac{1}{4f^2}\left(\frac{\partial f_0}{\partial G_T}\right)^2 S_{G_T}^{1/f} \Rightarrow \\ \mathscr{L}(f) &= \frac{1}{4f^2}\left(\frac{\partial f_0}{\partial G_T}\right)^2 \left(\beta_{M_N}^2 S_{M_N}^{1/f} + \beta_{M_P}^2 S_{M_P}^{1/f}\right) \end{aligned} \tag{3.21}$$

where $S_{f_0}^{1/f}$ is the spectral density of flicker noise, while $S_{M_N}^{1/f}$ and $S_{M_P}^{1/f}$ are the spectral noise density contributions of transistors M_N and M_P, respectively; G_T is the conductance of the tunable resistor; and f_0 is the oscillation frequency from (3.15). Thus, $\partial f_0/\partial G_T$ can be easily obtained:

$$\frac{\partial f_0}{\partial G_T} = \frac{(g_{ds_n} + g_{ds_p} - g_{m_p})(g_{ds_n} + g_{ds_p})}{\left(G_T + g_{ds_n} + g_{ds_p}\right)^2} \frac{\tan(\pi/N)}{2\pi C_L} \tag{3.22}$$

On the other hand, the spectral noise density of the tunable resistor must take into account the contribution of transistors M_N and M_P. In the literature, carrier density fluctuation is consider to be the dominant component of flicker noise in NMOS transistors. This is caused by carrier trapping and emission at the interface between silicon and insulator at the MOS gate [109, 110, 111, 112, 113].

In [111], the McWhorter model of spectral noise density for NMOS transistors in linear region is defined as,

$$S_{M_N}^{1/f} = \frac{K_n}{C_{ox}^2 (WL)_{M_N} f^{\eta}} \tag{3.23}$$

where K_n is a process dependent constant, generally lower for cleaner process and its magnitude can vary from 10^{-27} to $10^{-25}\,\mathrm{C}^2/\mathrm{m}^2$; C_{ox} is the gate oxide capacitance per unit of area, WL is the transistor area, and η is a technological parameter that varies in a range between 0.7 and 1.2, but it could be consider 1 if trap density is uniform [110, 111]. Carrier mobility fluctuations has been historically attributed to be responsible of flicker noise in PMOS transistors [110]. In [111, 110], the Hooge model of spectral noise density for PMOS transistors is defined as,

$$S_{M_P}^{1/f} = \frac{K_p(V_{SG})}{C_{ox} (WL)_{M_P} f} \tag{3.24}$$

where $K_p(V_{SG})$ is a process and bias dependent constant, typical values range from from 6×10^{-26} to $2 \times 10^{-23}\,\mathrm{V}^2\,\mathrm{F}$ [110]. It can be observed that the phase noise of this stage depends on the strength of the positive feedback term,

$$S_{f_0}^{1/f} = \left[\frac{(g_{ds_n} + g_{ds_p} - g_{m_p})(g_{ds_n} + g_{ds_p})}{2\pi C_L \left(G_T + g_{ds_n} + g_{ds_p} \right)^2} \tan\left(\frac{\pi}{N}\right) \right]^2 S_{G_T}^{1/f} \tag{3.25}$$

As expression (3.25) shows, phase noise due to flicker noise can be reduced by increasing the number of delay stages. Moreover, the phase noise is inversely proportional to C_L^2 and $S_{G_T}^{1/f}$ is inversely proportional to the length of M_N and M_P transistors, however, any increment in these parameters reduces the oscillation frequency.

Pseudo-Differential Stage with Cross-Coupled Inverters

Again, phase noise due to flicker noise is defined by (3.21) since the tunable resistor is equivalent for both stages. Hence, the difference between the spectral density of both stages is given by $\partial f_0 / \partial G_T$,

$$\frac{\partial f_0}{\partial G_T} = \frac{\left(g_{ds_n} + g_{ds_p} \right)_z^2}{\left[G_T + (g_{ds_n} + g_{ds_p})_z \right]^2} \frac{\tan(\pi/N)}{2\pi C_L} \tag{3.26}$$

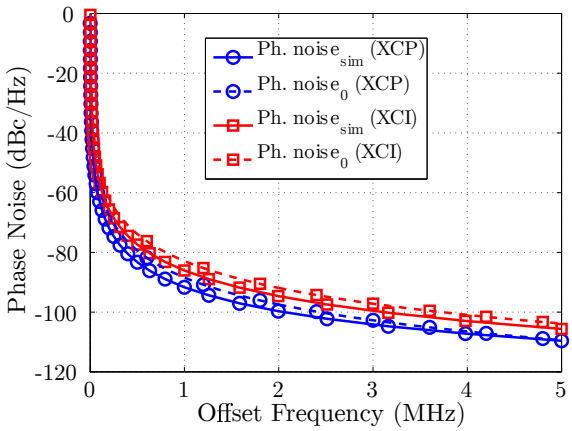


Figure 3.17: Phase noise as a function of offset frequency. Circle markers represents the data of XCP stage, and square markers represents the data of XCI stage. Solid lines represent the simulated data and dashed lines represent the small-signal model.

Finally, it can be observed that the spectral density of flicker noise, and hence the phase noise, of this stage do not depends on the strength of the positive feedback term,

$$S_{f_0}^{1/f} = \left\{ \frac{\left(g_{ds_n} + g_{ds_p}\right)_z^2}{2\pi C_L \left[G_T + (g_{ds_n} + g_{ds_p})_z\right]^2} \tan\left(\frac{\pi}{N}\right) \right\}^2 S_{G_T}^{1/f} \tag{3.27}$$

Expressions (3.27) and (3.25) differ only on the presence of positive feedback term, this suggest that XCP rings have smaller phase noise that XCI rings.

Simulation Results

Phase noise as a function of offset frequency is shown in figure 3.17 for the XCP and XCI stages. The phase noise of the XCP and XCI stages is -91 dBc/Hz and -86 dBc/Hz at 1 MHz, respectively. As predicted by the analytical model, the phase noise of the XCP stage is smaller than the phase noise of XCI stage, in this case, 5 dB/Hz at 1 MHz. The parameters of (3.23) and (3.24) are: $K_p = 10^{-26}\,\mathrm{V^2F}$, $K_n = 3\times10^{-28}\,\mathrm{C^2/m^2}$, $\mu_n = 0.03695\,\mathrm{m^2s/V}$, $\mu_p = 0.010591\,\mathrm{m^2s/V}$, $C_{ox} = 10\times10^{-3}\,\mathrm{F/m^2}$ and $\eta = 1$. On the other hand, the parameters of (3.25) are: $g_{ds_n} + g_{ds_p} = 180\,\mu$S, $g_{m_p} = 38\,\mu$S, $G_T = 160\,\mu$S, and $C_L = 14.5\,$fF. The parameters of (3.27) are: $(g_{ds_n} + g_{ds_p})_z = 280\,\mu$S, $G_T = 160\,\mu$S, and $C_L = 14.5\,$fF.

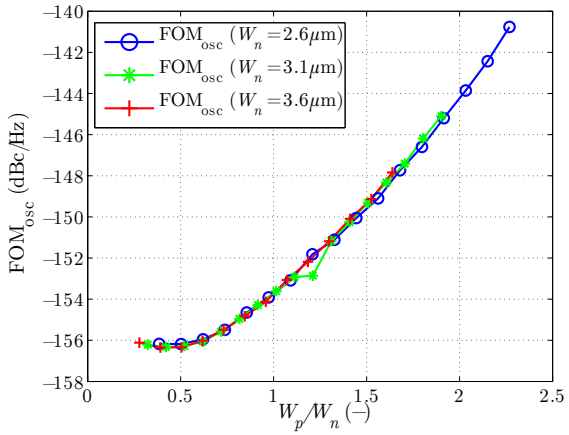


Figure 3.18: FOM$_{\text{osc}}$ of a ring with 3 XCP stages as a function of W_p/W_n with $V_{Bias} = 0$ V.

3.3.3 Application

Despite the fact that both stages are used indiscriminately for the same applications in the literature, the analysis done so far shows that the XCP stage has lower phase noise and the XCI stage has higher oscillation frequency. Hence, it is mandatory to study the FOM of each stage for each application. For example, the principal FOM of oscillators is defined by [99],

$$\text{FOM}_{\text{osc}} = \mathscr{L} - 20\log_{10}\left(\frac{f_0}{f_{off}}\right) + 10\log_{10}\left(\frac{P}{1\text{mW}}\right) \tag{3.28}$$

where $\mathscr{L}$ is the single side band phase noise power spectral density, f_0 is the oscillation frequency, f_{off} is the offset frequency at which the phase noise is measured, and P is the power consumption in milliwatts. In figure 3.18, it is shown that the FOM$_{\text{osc}}$ of the XCP stage does not depend on the size of NMOS transistor. Moreover, a minimum is reached at the ratio $W_p/W_n = 0.5$.

On the contrary, the FOM$_{\text{osc}}$ of the XCI stage does not have any minimum, and its magnitude depends on the size of W_{nz}, showing the same behavior as the oscillation frequency. Also, the smallest value of FOM$_{\text{osc}}$ in this oscillator is similar to the obtained with the XCP stage, however, this can only be achieve increasing W_{nz}. This results suggest that XCP stages are more robust, despite having lower oscillation frequency.

VCRO are commonly used in many TDCs architectures. The basic FOM for TDCs represents the amount of energy consumed during a conversion [114],

$$\text{FOM}_{\text{TDC}} = \frac{P}{2^{\text{ENOB}} f_s} \tag{3.29}$$

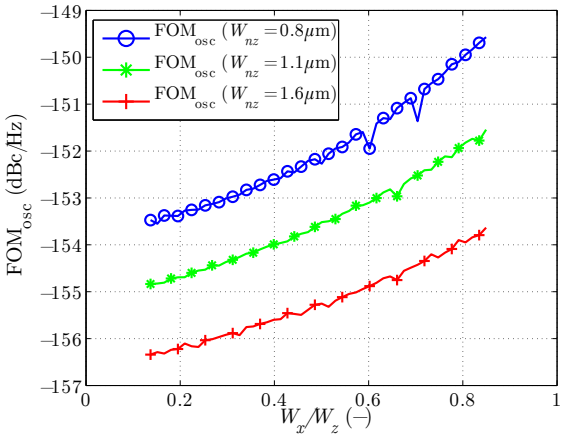


Figure 3.19: FOM$_{\text{osc}}$ of a ring with 3 XCI stages as a function of W_x/W_z with $V_{Bias} = 0$V and $W_{pz} = 3.6$µm.

where P is the power consumption, f_s is the sampling frequency of the converter, and effective number of bits (ENOB) is the effective number of bits. The ENOB can be approximated as (3.30) if the effect of systematic deviations are negligible, and only random contributions are consider [115],

$$\text{ENOB} = \log_2\left(\frac{T_{fs}}{\sigma_{T_0}\sqrt{12}}\right) \tag{3.30}$$

where $T_{fs} = 1/f_s$ is the full-scale time of the converter, fixed to 20 ns for every converter. On the other hand, the period jitter, σ_{T_0}, can be obtained from the phase noise [109],

$$\sigma_{T_0} = \frac{f_{off}}{f_0}\sqrt{\frac{\mathscr{L}}{f_0}} \tag{3.31}$$

where f_{off} is the offset frequency at which the phase noise is measured, f_0 is the oscillation frequency, and $\mathscr{L}$ is the single side band phase noise power spectral density. figure 3.20 and 3.21 illustrate the amount of energy consumed in a conversion (FOM$_{\text{TDC}}$) as a function of the transistor ratios for the XCP and XCI stages. It can be observed that the FOM$_{\text{TDC}}$ of both stages is reduced as the positive feed-back term is reduced. While the XCP stage shows a drastic reduction from 3 pJ per conversion down to 0.5 pJ per conversion, the XCI stage provides the lowest values (0.3 pJ per conversion). As a result, the XCI stage provides more efficient conversions than the XCP stage.

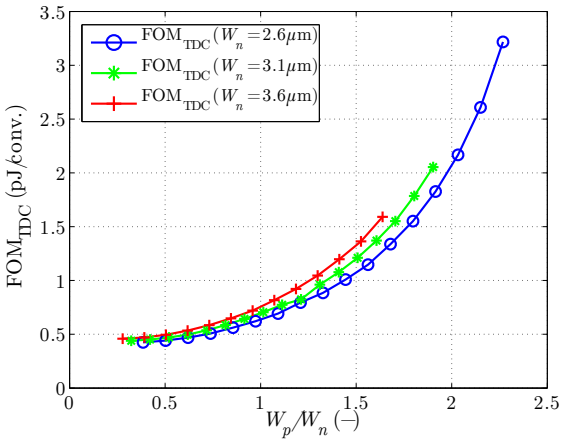


Figure 3.20: FOM$_{TDC}$ of a ring with 3 XCP stages as a function of W_p/W_n with V_{Bias} = 0 V.

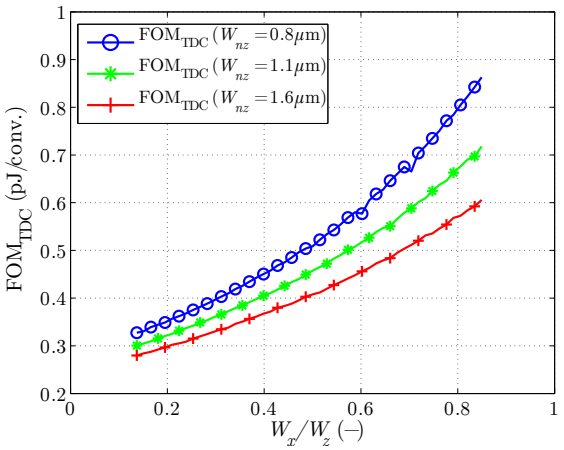


Figure 3.21: FOM$_{TDC}$ of a ring with 3 XCI stages as a function of W_x/W_z with V_{Bias} = 0 V and W_{pz} = 3.6 µm.

3.3.4 Evaluation of Results

The previous sections concentrates on the modeling and design optimization of two representative VCROs based on pseudo-differential delay cells. Although the outputs of this kind of oscillators are switching from rail-to-rail, small-signal analysis remains a simple and quite accurate tool to compute a first order approximation of the oscillation frequency and noise. This is useful to obtain a fast optimization of the design parameters.

The small-signal model provides good accuracy when the number of stages

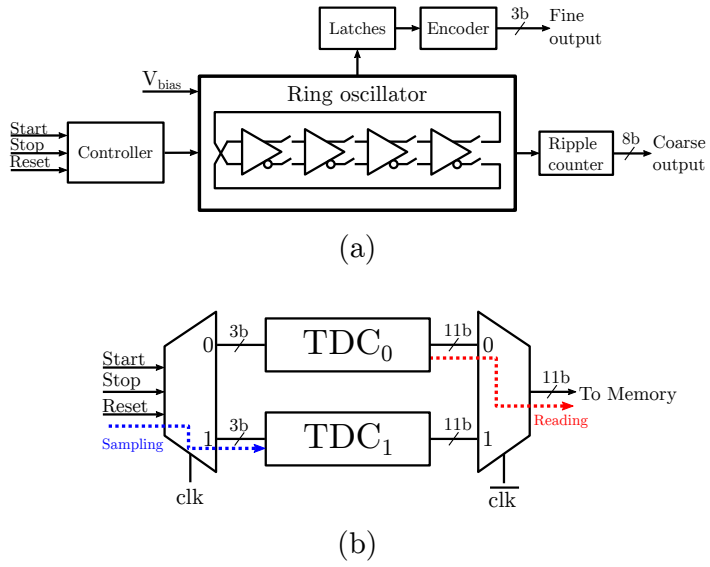


Figure 3.22: (a) Block diagram of TDC based on VCRO with 4 delay stages and (b) two TDC working interleaved.

is small, if the time spent with saturated outputs is minimum. Additionally, this model highlights very well the contribution of each transistor to the oscillation frequency, making the comparison between stages quite straightforward and helping the designer to identify the principal trade-offs. However, this model can not predict the oscillation frequency when the positive feedback term is big or the number of stages is large.

Analytical expressions agree with simulation results, showing that a ring oscillator made of pseudo-differential stage with cross-coupled inverters have higher oscillation frequency. At the same time, analytical expressions predict smaller phase noise in ring oscillators composed of pseudo-differential stage with cross-coupled PMOS and simulation results confirm it.

Finally, FOMs were used to study the best application for each stage. Both stages show similar performance as oscillator, since they seem to have the same phase noise figure at the best design choice. On the other hand, XCI stage is better suited for TDCs, when power consumption is considered.

3.4 Time-to-Digital Converter Implementation

The study of delay stages has provided an insight into the relations of time resolution and phase noise with the basic design parameters. This approximation allows to find the design point at which the oscillation frequency is maximum, while keeping a reasonable value of power consumption, without having to deal

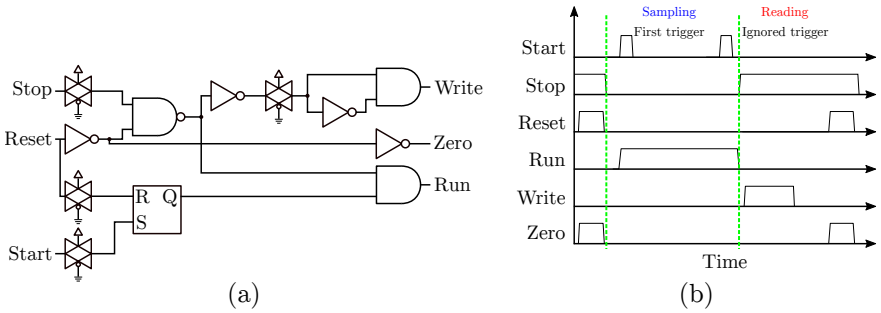


Figure 3.23: (a) block diagram and (b) time diagram of controller.

with cumbersome expressions. Figure 3.22-(a) illustrates a simplified block diagram of a TDC based on VCRO. The building blocks that compose the converter are responsible to start and stop the oscillations, sample the oscillator phases, encode the pseudo-thermometric code[1] of the oscillator phases into binary and count the number of cycles that the ring has done during oscillation. In the following sections these circuits will be presented and described, putting emphasis on its practical implementation.

One of the peculiarities of TDC designed for PET is that the trigger signal is completely asynchronous, since the arrival time of gamma photons is unknown. This means that the TDC must cover the whole temporal window. Therefore, architectures with low conversion time are preferred, since the number of TDCs working interleaved is reduced and, hence, the occupation area. The VCRO architecture have a very small conversion time, allowing to cover the whole clock cycle with only 2 interleaved TDCs. Ideally, this implies that while a TDC is sampling, the other is being read, see figure 3.22-(b). Nevertheless, there is always a small misalignment of clock signals that leads to a dead time in the transition between sampling and reading.

3.4.1 Control Logic

As figure 3.22 illustrates, the control block is responsible for generating the signals that allow the normal operation of the converter from the input signals start, stop and reset. The output signals of this block must put the oscillator phases to a known value, generate the turn on signal to start the oscillations with the first trigger and generate the signal to read the oscillator phases:

- Run: the ring oscillates while this signal is active, this means, that the

[1] In a thermometric code, the number "N" is encoded as a string of "N" '1's. Thus, the information is encoded at the transition between '1's and '0's, e.g. the numbers 5 and 6 are {11111000} and {11111100}, respectively. On the contrary, in the pseudo-thermometric of this ring oscillator, '1's and '0's are alternated and the information is encoded with a string '11' or '00', e.g. the numbers 5 and 6 are {10101101} and {10101001}, respectively.

pulse width, T_{sample}, is given by the rising edge of the start and stop signal. To guarantee that the measured interval correspond to the arrival time of the first photon, the circuit must only respond to the first start pulse.

- Write: when this signal is active, the latches sample the oscillator phases that provide the LSB of the conversion. This signal is generated once the oscillations are stopped. Since the registers are active by level, this signal must have the proper width in all corners.

- Zero: this signal is needed to set to a know value the latches, the ripple counter and internal phases of the ring. Doing so, the oscillations start always from the same point. Despite this alternative is the simplest, there are circuits in the literature that implements first order noise shaping by using the stopping point of the previous conversion as the starting point of the new conversion [83].

The circuit of figure 3.23-(a) implements the previous specifications, as the time diagram of figure 3.23-(b) shows, the inputs of the RS bistable is never at the high level at the same time during the normal operation, because the demuxiplier of figure 3.22-(b) prevents against it. Also, some transmission gates have been added to compensate the delays of the start and stop signals.

3.4.2 Ring Oscillator

As it was proved previously, the TDC designed with XCI delay stages perform better than the XCP counterpart. In order to have a design as compact as possible, a ring with 4 delay stages is the common choice among designer [89, 116, 105]. This means that the number of phases at which the oscillator can be stopped is 8. As it is expected, these phases can be encoded in 3 bits. In this topology, the phase noise and occupied area of the ring are affected by the number of delay stages in the ring, but the temporal resolution is not modified, because the stage delay is kept constant. Keeping that in mind, if it were necessary to reduce the converter power consumption and the occupied area is not a restriction, a ring with 8 delay stages could be an alternative. Since the oscillation frequency is halved, the power consumption of the ripple counter is reduced considerably.

Figure 3.24 presents a ring oscillator composed of 4 delay stages with tunable resistors, reset switches and output inverters. The input of the latches are connected to f_1-f_4 and $\overline{f_1}$-$\overline{f_4}$, while the input of the ripple counter is connected to f_4 and $\overline{f_4}$. For that reason, these inverters do not have an enable input, since they must let pass the value of f_4 and $\overline{f_4}$ during oscillations. The main purpose of the other inverters is to make the load capacitance of each stage equal, this is a common practice to reduce non-linearities between stages.

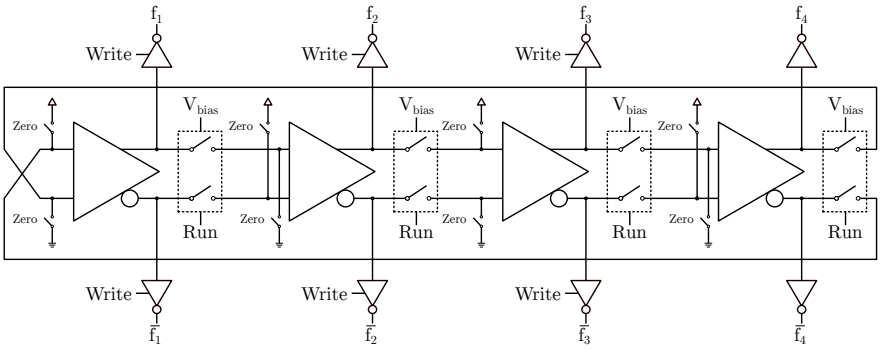


Figure 3.24: Voltage-controlled ring oscillator with 4 delay stages.

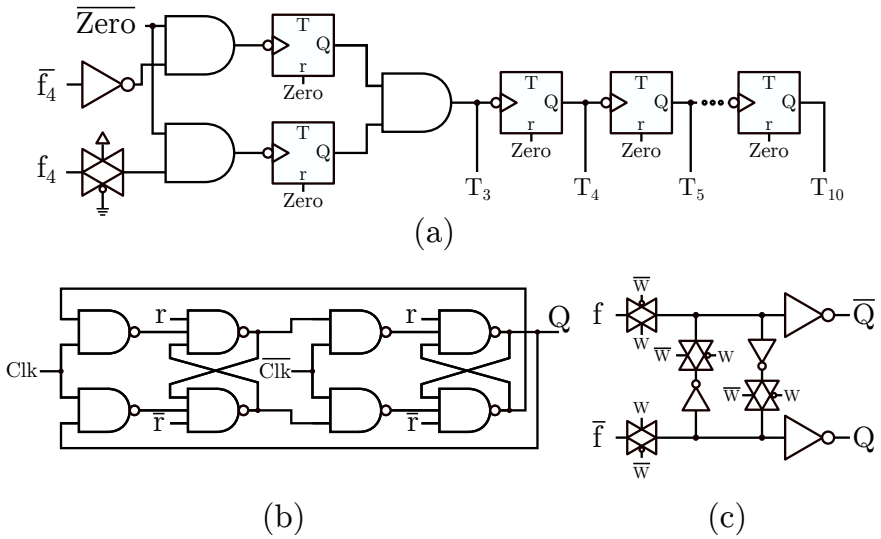


Figure 3.25: (a) Ripple counter based on toggle flip-flops, (b) toggle flip-flop implemented with NAND gates and (c) cross-coupled latch.

3.4.3 Read-out Logic

The main function of these blocks is to provide the corresponding binary value associated to the number of oscillations and the stopped phase. The first functionality can be implemented with a ripple counter connected to the last oscillator phase (f_4). Hence, when f_4 has a falling transition, the counter increases its value. In practice, this kind of implementations are prone to suffer from glitches when the oscillator is stopped during a transition of f_4. This entails the formation of a non-monotonicity in the static transfer function of the converter. The main reasons why these glitches appear are the metastability of delay stages, because of its positive feed-back, and the non-symmetric switch-off of the tun-

Table 3.2: Phases of a ring oscillator with 4 stages.

f_1	f_2	f_3	f_4	T_2	T_1	T_0
1	0	1	0	0	0	0
0	0	1	0	0	0	1
0	1	1	0	0	1	0
0	1	0	0	0	1	1
0	1	0	1	1	0	0
1	1	0	1	1	0	1
1	0	0	1	1	1	0
1	0	1	1	1	1	1

able resistor. The reader may have noticed that, as figure 3.9-(c) shows, during the switching off transition, the NMOS transistor is disabled before the PMOS transistor. In order to reduce the impact of these glitches and to assure that the oscillation count is increased when a real transition occurs, two flip-flops sample the phases f_4 and $\overline{f_4}$ and their outputs are compared with an AND gate, as figure 3.25-(a) illustrates. Doing so, the LSB of the counter is only increased when the transition of the two signals happens.

The principal drawback of this solution is the power consumption increment, due to the additional flip-flop working at the oscillation frequency. Analytically, the power consumption of the ripple counter can be approximated as,

$$\begin{aligned} P_c &= P_s + P_d \\ P_s &= V_{dd} I_{leak} \\ P_d &\approx f_{osc} C_{eq} V_{dd}^2 \left(2 + \sum_{i=3}^{n=10} \frac{1}{2^{i-2}} \right) \end{aligned} \tag{3.32}$$

where P_c is the total power consumption of the ripple counter, P_s is the static power consumption due to leakage, P_d is the dynamic power consumption, V_{dd} is the voltage supply, I_{leak} is the current leakage, f_{osc} is the oscillation frequency, C_{eq} is the equivalent input capacitance and i is the bit index. The factor 2 inside the braces represents the two flip-flops. As it is expected, the topology of the flip-flop will determine the power consumption by means of C_{eq}. figure 3.25-(b) represents the schematic of a NAND flip-flop, this topology has been chosen because of two reason: it is a static memory, therefore, the stored value is not lost and no minimum sampling frequency is required, f_s; and its maximum operation frequency is high enough to deal with the output of the oscillator. The main drawback of this topology is its large area occupation compared to other alternatives.

Regarding the phase sampling, two circuits are needed: regenerative latches like the one shown in figure 3.25-(c) and an encoder to translate the pseudo-

Table 3.3: Post-layout simulation results.

Parameter	Value	Units
Area	2925	μm^2
Ring oscillator power	0.8	mW
Ripple counter power	0.6	mW
Total power	1.4	mW
Maximum range	185	ns
Maximum f_s	50	MHz
Minimum resolution	90	ps

thermometric code of the phases into binary. To find the encoder expression, the different phase permutations during oscillations must be studied. The truth table 3.2 represents all possible permutations of the phases and their encoded value. It can be observed that between consecutive permutations, only one bit is changed, similar to what happens in Gray code. Using the truth table, the expressions for the of the converter output can be easily obtained, as (3.33) represents.

$$T_0 = f_1\, f_2\, \overline{f_3}\, f_4 + \overline{f_1}\, \overline{f_2}\, f_3\, \overline{f_4} + f_1\, \overline{f_2}\, f_3\, f_4 + \overline{f_1}\, f_2\, \overline{f_3}\, \overline{f_4} \tag{3.33a}$$

$$T_1 = f_1\, \overline{f_2}\, f_4 + \overline{f_1}\, f_2\, \overline{f_4} \tag{3.33b}$$

$$T_2 = f_4 \tag{3.33c}$$

The simplicity of these expressions allows to implement them with CMOS, transmission gate or pseudo-NMOS logic, to name a few. These expressions were implemented with pseudo-NMOS logic, because of its small footprint.

3.4.4 Physical Design and Simulation Results

Post-layout simulation results are summarized in table 3.3, it can be see that the occupation area is bigger than in [89], mainly because of the empty space and the large area of the ripple counter. Moreover, the power consumption is over 1 mW, while the ring is oscillating at maximum frequency. Also, the minimum time resolution is below 100 ps. However, the final accuracy of the converter will be worse, because there is no phase locked loop (PLL) to compensate PVT variations. Future designs will address this issue. All in all, it is expected a degradation of these parameters during the chip measurements.

Figure 3.26 illustrates the layout of the designed converter. It can be see that the ripple counter is the biggest block followed by the ring oscillator. Although the blocks themselves are compact, the floorplan shows that there is still room for future improvements in terms of area occupation.

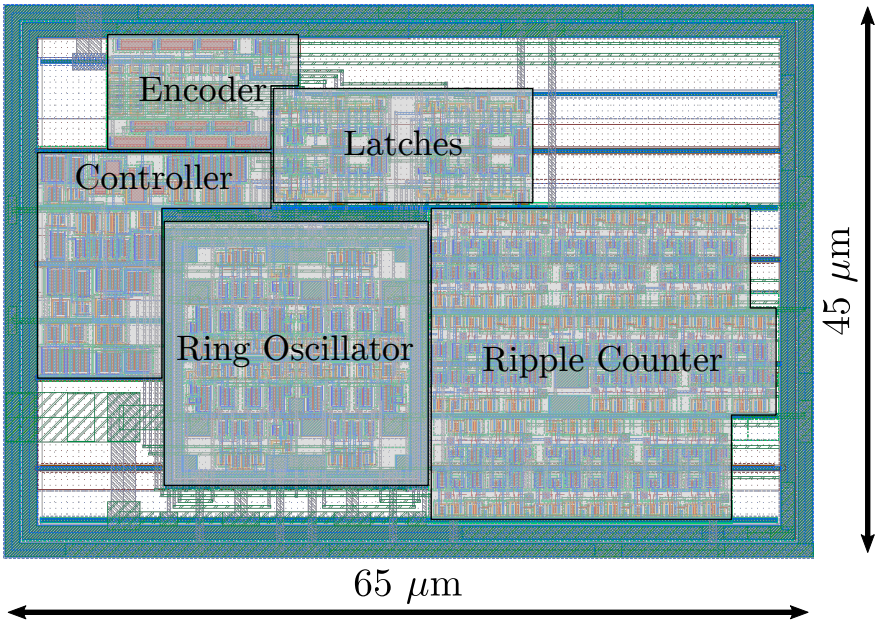


Figure 3.26: TDC layout.

CHAPTER 4

ARCHITECTURAL EXPLORATION

As it was discussed previously, SiPMs are a promising solid-state alternative to substitute PMTs in a large number of applications. Because of their lower power consumption, lower bias voltage, smaller size, higher spatial resolution and magnetic field compatibility, to name a few advantages. However, the design of such sensors is not trivial, because there are some trade-offs that make impossible a design to be neither straightforward nor optimal for every application, like sensitivity versus time resolution, sensitivity versus linearity or FF versus cross-talk. Therefore, it is necessary to choose carefully the proper parameters to be optimized in each case. From the analog designer point of view, this optimization problem is bounded to geometric constrains, since the modification of the doping profile and junction depth is beyond his/her reach. As a first order approximation, it can be said that the sensor linearity is a function of SPAD/mm^2, cross-talk is a function of the distance between SPADs, while FF is a function of SPAD size. Based on this hypothesis, the optimization is only possible if the required data is available during the first steps of the design and the analytical descriptions of the sensor and the environment are accurate enough.

The exploration of the design space is done by selecting the most important FOM for a given application and, then, optimizing this FOM respect to different design variables. This procedure avoids the selection of sub-optimal architectures and helps to justify design decision by highlighting the impact of each variable on the sensor performance.

The aim of this chapter is to provide an analytical study of the design space of digital silicon photomultiplier (d-SiPM) for PET-ToF. To do so, the basic pixel architecture is described in section 4.1, along with the functional blocks that satisfy the desired features for PET-ToF. Section 4.2 describes the study of the

design space that determines the optimum architecture in term of FOM. Two approaches are consider there: firstly, the SPAD size is treated as another design variable; and, secondly, the SPAD size is kept constant. Finally, the proposed architecture is described in detail in section 4.3.

4.1 Pixel Definition

As it was explained in chapter 1, the d-SiPMs designed for PET-ToF applications detects the light produced by a scintillator crystal when a gamma photons interacts with it. This means that the detector block should be composed of an array of d-SiPM coupled to a monolithic crystal, generally LSO or LYSO, to provide multiple time stamps and DOI correction. Since monolithic crystals are used, the generated optical photons during the event are distributed all over the array and, as a result, the amount of incident photons over each sensor is relatively small. Therefore, special care during the design must be taken to extract all the information possible from these few photons: arrival time of the incident particle, energy transfer to the crystal and the first interaction point inside the crystal.

Based on these considerations, the pixel functionality can be decomposed into three main parts: a block to process scintillation events and discard noise, implemented by a FSM; a block to determine the arrival time of gamma photons at the scintillator, which is a TDC; and a block to perform the photon integration, this is achieved with SPADs connected to a rippler counter through a pull-up OR. As it was discussed in chapter 1, the first interaction point inside the crystal can be obtained by studying the light profile over the whole sensor block, therefore, the data from the whole array of SiPMs is needed. figure 4.1 shows a simplified diagram of the d-SiPM that contains all the functional blocks, note that between sub-pixels and the FSM there is an adder tree that sets the hold-on and set-up restrictions, i.e. more sub-pixels means more branches and, therefore, longer delay.

While the FSM and TDC are considered as a black box during this analysis, that is, only their footprint area are consider, the variables of this analysis are: the SPAD size, the number of sub-pixels per pixels and the number of SPADs per micro-cell. The SPAD size determines the FF and the number of SPAD per pixel. The number of sub-pixels have a direct impact in the linearity of the sensor, because each sub-pixel counts the detected photons to later send them through the adder tree (see figure 4.1). At the same time, the number of SPADs per micro-cell is a determinant for FF and spatial compression [81], because the SPADs in the same micro-cell share some logic and the dead time.

Before starting the architectural exploration, it is needed to summarize all the constants that plays role during the design. Some of these parameters, like the pixel area, the maximum number of incident photons per unit of area and the scintillator time constant, have been presented during the definition of

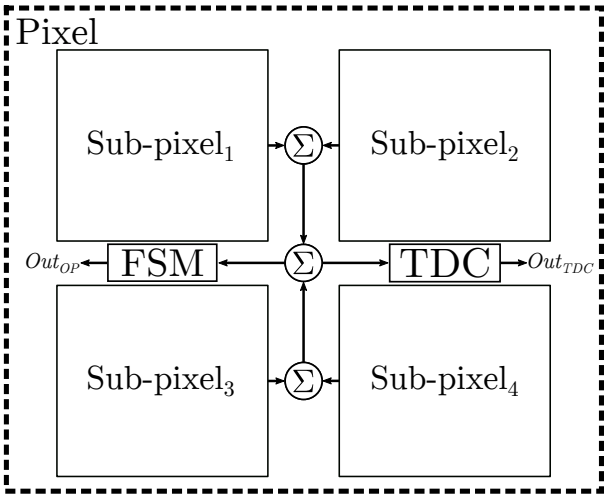


Figure 4.1: Simplified diagram of a digital SiPM pixel with an adder tree of 2 levels and the SPADs clustered into 4 sub-pixels. Each pixel provides both the number of counted photons (Out_{OP}) and the arrival time of the incident particle (Out_{TDC}).

the sensor environment in chapter 1. Other parameters were obtained from the full-custom library blocks described in the previous chapters, like the area of the FSM, TDC and AQRC. Also, post-layout simulations haven been carry out to estimate the time performance of these library blocks in the worst case scenario, i.e. the clock period is set by the maximum conversion time of the TDC and the width of the monostable pulse is determined by the bandwidth of the TDC trigger path. The last set of values, like DCR and PDP, were obtained from measurements over the chip, since they can not be estimated indirectly. With the constants of table 4.1 and the number of sub-pixels per pixels and the number of SPADs per micro-cell as variables, the pixel performance can be analytically estimated by the expression obtained in the next sections.

4.2 Figure of Merit Definition and Optimization

Among the SiPM fabricated nowadays, two different design paradigms can be identified: low and bright light conditions. This forces to choose different FOMs for each light condition. On the one hand, the PDE is chosen as a principal FOM for the SiPMs working in low light conditions, like PET and FLIM. This leads to SPAD pitch between 30 – 60 µm and fill factors ranging 50 – 70 % [117, 43]. The use of PDE as a reference in PET-ToF systems is mainly motivated by the fact that gamma photon arrival time jitter is strongly dependent on the number of detected photons, rather than being dependent on the SPAD jitter or the TDC single shot precision, due to lower photon statistics [118, 119].

Table 4.1: Parameter values used for the architectural exploration. These values are obtained from our own library blocks, simulations and measurements.

	Parameter	Value	Units	Description
Geometric	A_{pxl}	1.69	mm^2	Pixel area
	A_{tdc}	6.5×10^{-3}	mm^2	area
	A_{fsm}	14×10^{-3}	mm^2	area
	A_{cnt}	2.6×10^{-3}	mm^2	Counter and adder tree area
	A_{aqrc}	90×10^{-6}	mm^2	Active quenching and reset circuit and area.
	A_{dcmm}	205×10^{-6}	mm^2	SPADs shared dead area (like monostable, reset...)
Luminous	PDP	30	%	Photo detection probability
	H_q	450	$1/mm^2$	Maximum incident photons per unit of area per event calculated in expression (1.9)
Temporal	T_{pulse}	800	ps	Monostable pulse width
	T_{clk}	20	ns	Clock period
	τ_{LSO}	40	ns	Time constant of LSO
	DCR	4×10^5	Hz/mm^2	Dark count ratio per SPAD active area
	T_{fa}	0.3–0.7	ns	Minimum and maximum full adder delay
	T_{path}	0.6	ns	Delay between full adders
	T_{skew}	2.0	ns	Skew clock time
	T_{set-up}	0.4	ns	Flip-flop set-up time
	$T_{hold-on}$	0.4	ns	Flip-flop hold-on time
	$T_{clk \to q}$	0.3–0.6	ns	Minimum and maximum flip-flop clock to data time

On the other hand, linearity is generally chosen as a reference FOM for SiPMs designed for bright light applications, like HEP. Linearity is linked to the density of SPADs per square millimetre $\left(\mathrm{SPAD/mm^2}\right)$ because the response of a SiPM, N_{count}, as a function of the incident photons, N_{ph}, can be simplified as:

$$N_{count} = N_{\mu c}\left[1 - \exp\left(-\frac{N_{ph}PDE}{N_{\mu c}}\right)\right] \approx N_{ph}PDE \quad \Leftrightarrow \quad N_{\mu c} \gg N_{ph}PDE \tag{4.1}$$

where the number of SPAD micro-cells, $N_{\mu c}$, is proportional to SPADs per square millimeter for a given SiPM size. Expression (4.1) can only be approximated to a linear function if and only if the number of SPADs is much higher than the number of detected photons ($N_{\mu c} \gg N_{ph}PDE$). With this in mind, the SPAD pitch of some designs have been push down to 5–20 µm allowing a maximum of 46 $\mathrm{kSPAD/mm^2}$ with a fill factor of 40 % [120]. However, such performance can

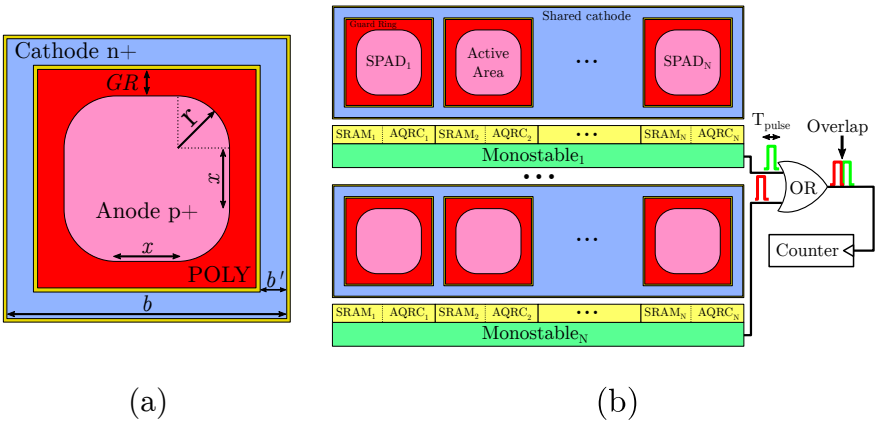


Figure 4.2: (a) SPAD parametric cell. The SPADs have a square shape with rounded corners, when $x > 0$. $r = 5\mu m$, $GR = 1\mu m$, $b = x + 2(GR + r + b') = x + 16.2\mu m$ and $b' = 2.1\mu m$. (b) Block diagram of sub-pixel. The SPADs into the same micro-cell can only generate one trigger during T_{clk}. Wide T_{pulse} can saturate the counter input (pull-up OR).

only be achieved in custom technologies optimized for analog silicon photomultiplier (a-SiPM), where, generally, in-chip signal processing is not possible, because transistors are not available.

During this architectural exploration, the sensitivity, the number of counted photons respect to the incident photons, is used as a reference FOM. The sensitivity is similar to the PDE, but it models the non-idealities of the compression techniques implemented by this architecture [81].

4.2.1 Variable SPAD size

The architectural exploration must start with a general approach that can be later particularized. Hence, the first step is to find the optimum SPAD size in terms of sensitivity. To do so, an analytical expression for sensitivity has to be obtained, derived with respect to the SPAD size and set it equal to zero. This approach provides the optimum SPAD size in terms of sensitivity for each design point, that is, for each combination of sub-pixels per pixel and SPADs per micro-cell, see figure 4.2. Then, the non-linearity and SNR of the pixels can be calculated to verify if the proposed architecture have an acceptable performance. This approach provides designers an useful insight of the design space, but it does not substitute them in the decision making process, because this analysis is mainly restricted to geometrical constrains and some key parameters, like jitter or power consumption, are not reflected on it.

The first step in the architectural exploration is to obtain the number of SPADs per pixel, N_{spad}, as a function of the number of sub-pixels per pixels,

M_{spl}, and the number of SPADs per micro-cell, $N_{\mu c}$. Thus, the number of SPADs per pixel can be calculated as the fraction between the pixel area not occupied by logic and the area occupied by one SPAD and its logic:

$$N_{spad}\left(M_{spl}, N_{\mu c}\right) = \left\lfloor \frac{A_{pxl} - (A_{tdc} + A_{fsm} + A_{cnt} M_{spl})}{A_{spad} + A_{aqrc} + \frac{A_{dcmm}}{N_{\mu c}}} \right\rfloor \tag{4.2}$$

where $A_{spad} = b^2 = \left[x + 2\left(GR + r + b'\right)\right]^2$ is the SPAD area, see figure 4.2-(a); A_{pxl} is the total pixel area; A_{tdc} is the TDC area; A_{fsm} is the FSM area; A_{cnt} is the area occupied by sub-pixel logic, like counters and multiplexers; A_{aqrc} is the area of the AQRC; and A_{dcmm} is the SPADs shared dead area, like monostable, reset circuits. Note that the parametric cell of the SPAD generates an square active area with rounded corners if $x > 0$. As it was explained in chapter 2, rounded corner are preferred over right corners, because they are associated with lower electric fields at the edges and, consequently, lower DCR due to premature breakdown [48]. The next step is to obtain the PDE of the pixel:

$$PDE\left(M_{spl}, N_{\mu c}\right) = \frac{A_{aspad} N_{spad}}{A_{pxl}} PDP \tag{4.3}$$

where $A_{aspad} = x^2 + 4rx + r^2\pi$ is the SPAD active area. As it was commented previously, an ideal SiPM has an output proportional to the number of incident photons. Nevertheless, since the number of micro-cells is limited in a real SiPM, its output experiments a saturation when the incident number of photons is high enough, in the order of $N_{\mu c} \sim N_{ph} PDE$ as expression (4.1) states. This, however, is only the first of the non-idealities that a digital SiPM has. The studied architecture presents some other limitations whose influence must be estimated. In [81], the probability of losing photons, pile-up, due to spatial and temporal compression is calculated with a Poisson distribution. The spatial compression is obtained when the monostable and reset logic is shared among different SPAD cells, see figure 4.2-(b). This means that only one photon can be counted by each SPADs micro-cell after an avalanche, because the dead time is shared among all the SPADs of the micro-cell. Here, to simplify the analysis, the SPAD dead time is set equal to the clock period, T_{clk}. The main advantage of this approach is the increase of fill factor, as a result of the fact that the area of common logic is shared among different SPADs. Hence, the probability of losing photons due to spatial compression (P_{sp}), i.e. two or more photons arrive at different SPADs from the same micro-cell during the scintillation build-up, from $t = 0$ to $t = T_{clk}$, can be expressed with a Poisson distribution [81]:

$$\begin{aligned} P_{sp}(\mu_{sp}; k > 1) &= 1 - P_{sp}(\mu_{sp}; k = 0) \\ &= 1 - \frac{\mu_{sp}^k}{k!} \exp\left(-\mu_{sp}\right) \end{aligned} \tag{4.4}$$

where k is the number of times the event occurs, that is, photon losses or pile-up per micro-cell; and μ_{sp} is the average number of events per sampling interval, i.e, the number of detected photons per micro-cell, from $t = 0$ to $t = T_{clk}$. As it was discussed in expression 1.4, the emission dynamics of a scintillation event are modeled by exponential decays. In the particular case of LSO and LYSO scintillators the emission is usually simplified with one time constant. The number of detected optical photons at a pixel per unity of area, OP_d, during the scintillation build-up, from $t = 0$ to $t = T_{clk}$ can be obtained by integrating the instantaneous photoemission described by expression (1.4) in chapter 1:

$$OP_d = PDE \cdot H_q \cdot \left[1 - \exp\left(-\frac{T_{clk}}{\tau_{LSO}}\right)\right] \tag{4.5}$$

where H_q is the maximum incident photons per unit of area per event calculated in expression (1.9) and τ_{LSO} is the scintillator time constant, in this case LSO. Finally, combining (4.4) and (4.5), the probability of losing photons due to spatial compression is:

$$P_{sp}(\mu_{sp}; k > 1) = 1 - \exp\left(-N_{\mu c} A_{aspad} OP_d\right) \tag{4.6}$$

From this expression it can also be derived that SiPMs composed of smaller SPADs have smaller losses due to spatial compression, i.e. larger linearity, because the active area of the micro-cell $\left(N_{\mu c} A_{aspad}\right)$ is reduced.

On the other hand, temporal compression is obtained when the monostable shrink the SPAD output pulses, making possible the counting of more pulses during the same readout interval. As there is not arbitration method to access the pull-up OR, the loss of information is expected due to pulse overlapping, see figure 4.2-(b). The probability of losing photons due to temporal compression, i.e. two or more monostable pulses are so close that the counter is unable to distinguish them, can be expressed with a Poisson distribution [81]:

$$\begin{aligned} P_{tp}(\mu_{tp}; k > 1) &= 1 - P_{tp}(\mu_{tp}; k = 0) \\ &= 1 - \frac{\mu_{tp}^k}{k!} e^{-\mu_{tp}} \\ &= 1 - \exp\left[-\frac{N_{spad} A_{aspad} OP_d T_{pulse}}{T_{clk} M_{spl}}\right] \end{aligned} \tag{4.7}$$

where k is the number of times the event occurs, that is, monostable pulses overlaps; and μ_{tp} is the average number of events per sampling interval, from $t = 0$ to $t = T_{clk}$. It is commonly attributed to the T_{pulse} to be the driving parameter of these losses, unfortunately its minimum value is limited by the resistive and capacitive parasitics of the trigger path. On the contrary, the fraction N_{spad}/M_{spl} is easier to modify, because it is independent of parasitics. Other approaches, like the presented in [121, 122], substitute the monostable and OR tree by a toggle register per micro-cell and a XOR tree. This way temporal compression is eliminated because the photons are encoded into the edges

of the XOR output and the dynamic range is accordingly increased. The main drawback of this approach is the larger occupation area of toggle registers per micro-cell and the XOR tree. Furthermore, the XOR tree is unable to preserve the arrival time of the first photon as OR tree does [122]. Finally, the sensitivity of the sensor, i.e. photon counting probability, can be calculated with (4.3), (4.6) and (4.7) as,

$$\begin{aligned} S(M_{spl}, N_{\mu c}) &= PDE(1-P_{sp})(1-P_{tp}) \\ &= A_{aspad} N_{spad} \frac{PDP}{A_{pxl}} \exp\left(-A_{aspad}^2 N_{spad} \alpha\right) \end{aligned} \tag{4.8}$$

where α is independent of the SPAD area:

$$\alpha = \frac{PDP}{A_{pxl}} H_q \left[1 - \exp\left(-\frac{T_{clk}}{\tau_{LSO}}\right)\right] \left(N_{\mu c} + \frac{N_{spad} T_{pulse}}{M_{spl} T_{clk}}\right) \tag{4.9}$$

Expression (4.9) is function of $N_{\mu c}$, M_{spl} and the SPAD size. Also, at its rightmost parenthesis, the contribution of spatial and temporal compression is shown. The term $N_{\mu c}$ determines the spatial compression, while the factor $N_{spad} T_{pulse} / M_{spl}$ sets the temporal compression. It can be see that that large values of the two components reduces the overall sensitivity. Also, the influence of temporal compression respect to spatial compression can be reduced if $N_{spad} T_{pulse} / M_{spl} \ll T_{clk}$. Therefore, since T_{clk} is determined by the TDC conversion time and T_{pulse} is limited by parasitics, the SPAD size, $N_{\mu c}$ and M_{spl} are the only free variable. In the following, the SPAD size, x, that provides the maximum sensitivity for each combination of M_{spl} and $N_{\mu c}$ must be calculated from $dS/dx = 0$. The analytical expression for dS/dx is:

$$\begin{aligned} \frac{dS}{dx} = \frac{PDP}{A_{pxl}} \Bigg\{ & \frac{d(A_{aspad} N_{spad})}{dx} \exp\left(-A_{aspad}^2 N_{spad} \alpha\right) \\ & + A_{aspad} N_{spad} \frac{d\left[\exp\left(-A_{aspad}^2 N_{spad} \alpha\right)\right]}{dx} \Bigg\} \end{aligned} \tag{4.10}$$

The last step is to set equal to zero expression (4.10). However, since (4.10) is a transcendental equation, the close-form expression to solve $dS/dx = 0$ may not exist. For this reason, numerical methods must be used to obtain the value of x that provides the best sensitivity for each combination of M_{spl} and $N_{\mu c}$. Figure 4.3 illustrates the optimum SPAD size in terms of sensitivity as a function of M_{spl} and $N_{\mu c}$. These pitches are similar to those offered by commercial products [38, 37, 42].

Finally, the SNR can be used to study the trade off between M_{spl} and $N_{\mu c}$. It can be approximated as follows if DCR (in Hz/mm^2) is considered to be proportional to the SPAD active area:

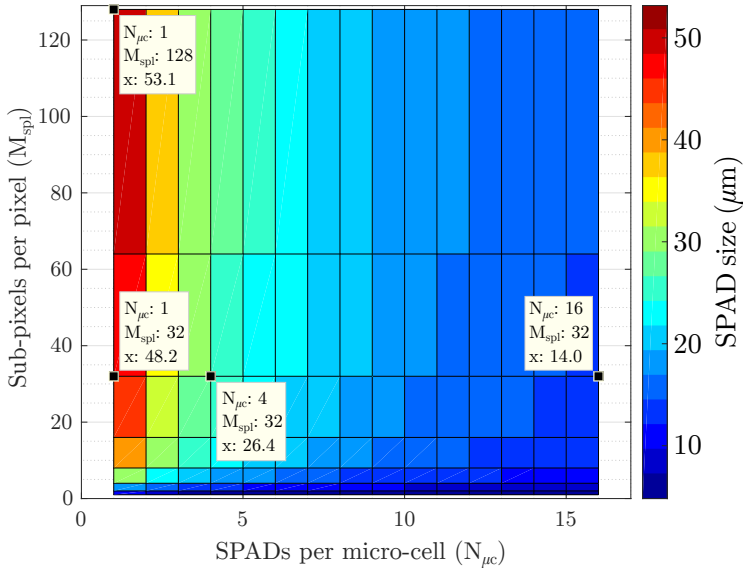


Figure 4.3: SPAD size, x, as a function of number of sub-pixels per pixel and number of SPADs per micro-cell.

$$
\begin{aligned}
SNR(M_{spl}, N_{\mu c}) &= \left(\frac{Signal}{Noise}\right)^2 = \left(\frac{OP_d A_{pxl} S}{T_{clk} N_{spad} A_{aspad} DCR}\right)^2 \\
&= \left(\frac{H_q \left[1 - \exp\left(-\frac{T_{clk}}{\tau_{LSO}}\right)\right] A_{pxl} S}{T_{clk} N_{spad} A_{aspad} DCR}\right)^2 \\
&= \left(\frac{H_q PDP \left[1 - \exp\left(-\frac{T_{clk}}{\tau_{LSO}}\right)\right]}{T_{clk} DCR}\right)^2 \exp\left(-2A_{aspad}^2 N_{spad} \alpha\right)
\end{aligned}
\tag{4.11}
$$

As figure 4.4 represents, the sensitivity has the lowest values when $M_{spl} < 16$, because the photon losses due to temporal compression are larger. However, when this number is above 16, the sensitivity reaches a plateau with the highest value at $M_{spl} = \{32, 64\}$ and $N_{\mu c} = 1$. These results show that architectures with shared monostables ($N_{\mu c} > 1$) have slightly lower sensitivity, because the sensor saturation is achieved earlier since the dead time is common among the SPADs connected to the same monostable.

The figure 4.5 illustrates the counted photons as a function of incident photons for a pixel with $N_{\mu c} = 1$. The dashed line represents the ideal response of a sensor (considering only the PDE), while solid lines show the response of

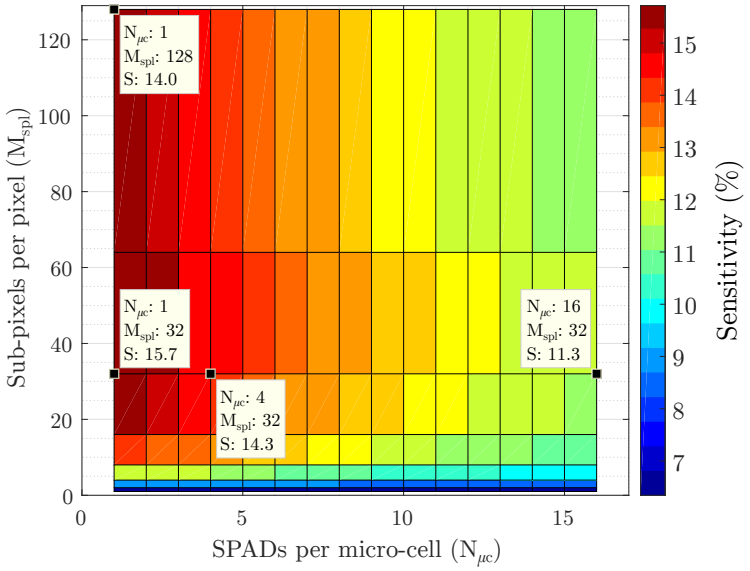


Figure 4.4: Sensitivity as a function of number of sub-pixels per pixel and number of SPADs per micro-cell for optimized size.

the sensor considering photon losses due to compression. In figure 4.6, the definition for the sensor non-linearity is provided. From this data, the non-linearity of the sensor can be calculated, as figure 4.7 summarizes. The lowest values of non-linearity can only be achieved if the parallelism of the pixel is maximum, i.e. SPADs do not share monostable and the number of sub-pixels is large.

Figure 4.8 illustrates the SNR as a function of $N_{\mu c}$ and M_{spl}. Although it may seem counter intuitive, the SNR does not have any maximum as the sensitivity does around $N_{\mu c} = 1$ and $M_{spl} = \{32, 64\}$. This is because the number of SPADs is smaller at the SNR maximum than at the sensitivity maximum (around 125 fewer SPADs). As a result, the small maximum is masked by noise reduction. It has to be noted that the DCR in this expression is not realistically modeled, since large SPADs have not only larger noise but their yield is worst (higher number of very noisy devices). Furthermore, in figure 4.8, it can be observed that the SNR is more or less constant, with less than 1 dB of variation, in the whole range of $N_{\mu c}$. The SNR is also quite constant respect to the M_{spl} when $M_{spl} > 16$.

At this point, the feasibility of these optimum architectures has to be verified with post-layout simulations, because the resistive and capacitive parasitics determine the violation of hold-on and set-up times at register-transfer level. Ideally, this should be done with digital synthesis tools. However, since all the cell where designed in the analog domain, an analytical approximation is

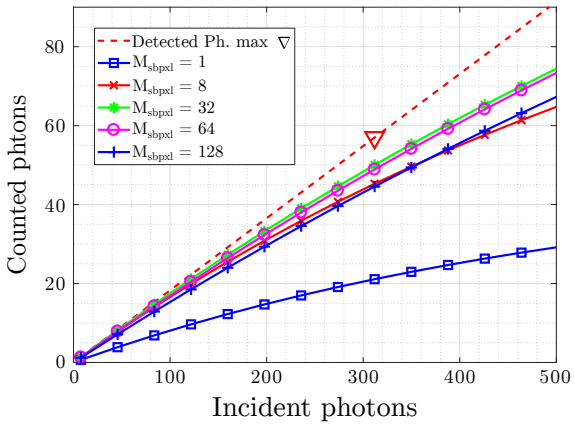


Figure 4.5: Counted photons as a functions of incident photons with 1 SPAD per micro-cell for optimized size. The dashed line represents the detected photons (ideal sensor response), the triangle is pointing the maximum number of incident photons in T_{clk}.

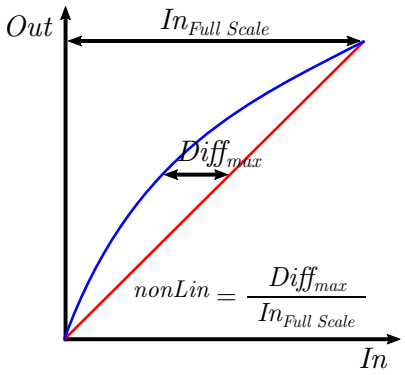


Figure 4.6: Graphic definition of non-linearity.

presented. In figure 4.9, the data propagation flow is shown. The number of accumulated photons at each sub-pixel must be sent through the adder tree to the FSM input. In order to avoid set-up and hold-on violations, the time needed for this data to be propagated, T_{logic}, determines the maximum and minimum number of sub-pixels for a given T_{clk}, respectively. This is because the number of sub-pixels determines the number of adders between the sub-pixels and the FSM. To avoid set-up violations, the data at the FSM input must be valid a time T_{set-up} before the clock edge. Thus, set-up condition is met if expression (4.12) is valid,

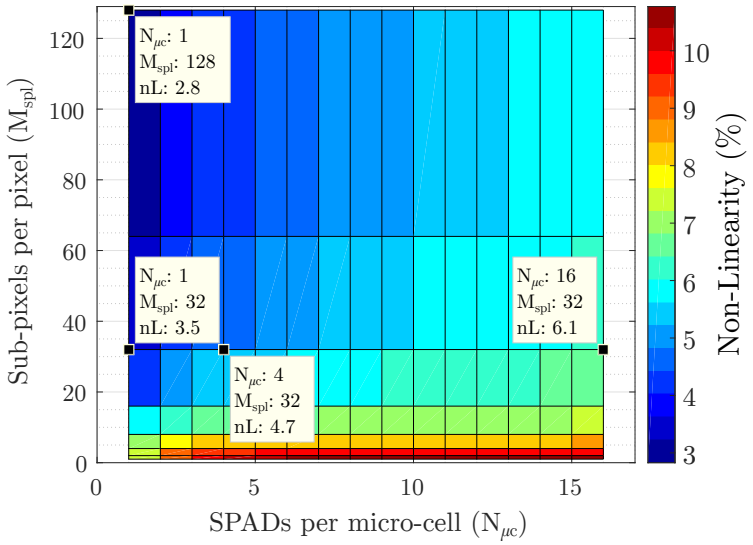


Figure 4.7: Pixel non-linearity for optimized size SPAD.

$$\begin{aligned} T_{clk} &\geqslant T_{skew} + T_{set-up} + T_{clk\rightarrow q} + T_{logic\ max} \\ T_{clk} &\geqslant T_{skew} + T_{set-up} + T_{clk\rightarrow q} + T_{fa}\left\{n + 2\left[\log_2\left(M_{spl}\right) - 1\right]\right\} + T_{path}\log_2\left(M_{spl}\right) \end{aligned} \tag{4.12}$$

where T$_{clk}$ is the clock period, T_{skew} is the delay in the clock path, T_{set-up} is the flip-flop set-up time, $T_{clk\rightarrow q}$ is the time needed for the flip-flop to show a valid output after the rising edge of the clock, $T_{logic\ max}$ is the maximum delay of the adder tree and it is divided into two component. T_{path} is the delay between two adder levels. $T_{fa}\left[n + 2\log_2\left(M_{spl}\right)\right]$ determines the longest delay path by determining the maximum number of full adders through which the data is propagated. The number of bits of the sub-pixel counter can be estimated taking into account the number of detected photons per sub-pixel,

$$n = \left\lceil \log_2 \left\{ \frac{H_q\ A_{pxl}\ PDE\ \left[1 - \exp\left(-\frac{T_{clk}}{\tau_{LSO}}\right)\right]}{M_{spl}} \right\} \right\rceil \tag{4.13}$$

The values of these parameters are calculated from our library blocks at the worst corner, resistive and capacitive parasitics are also included. On the other hand, to avoid hold-on violations, the data must not change a time $T_{hold-on}$ after the clock edge. The hold-on condition is met if expression (4.14) is valid,

$$\begin{aligned} T_{clk\rightarrow q} + T_{logic\ min} - T_{skew} - T_{hold-on} &\geqslant 0 \\ T_{clk\rightarrow q} + \left(T_{fa} + T_{path}\right)\log_2\left(M_{spl}\right) &\geqslant T_{skew} + T_{hold-on} \end{aligned} \tag{4.14}$$

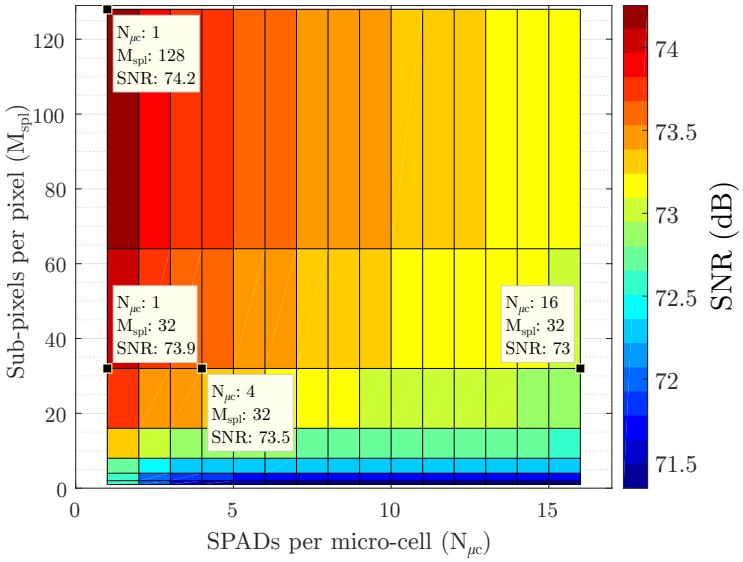


Figure 4.8: Signal to noise ratio as a function of number of sub-pixels per pixel and number of SPADs per micro-cell for optimized size.

where $T_{logic\ min}$ is the minimum delay of the adder tree and it is function of M_{spl}. Table 4.2 summarizes the pixel configurations that violate hold-on and set-up conditions. Considering that the clock period is fixed at 20 ns, due to the TDC conversion time, the number of sub-pixels per pixels must be higher than 4 to avoid hold-on violations. In terms of set-up, the number of sub-pixels could be 64 or 128. However, these results are too close to the clock period and should be taken with caution since they are approximations.

Table 4.2: Set-up and hold-on conditions for $N_{\mu c} = 1$.

M_{spl}	Set-up (ns)	Hold-on (ns)
1	$T_{clk} \geqslant 6.5$	$0.3 \ngeqslant 2.2$
2	$T_{clk} \geqslant 8.5$	$1.2 \ngeqslant 2.2$
4	$T_{clk} \geqslant 9.8$	$2.1 \ngeqslant 2.2$
8	$T_{clk} \geqslant 11.1$	$3.0 \geqslant 2.2$
16	$T_{clk} \geqslant 13.1$	$3.9 \geqslant 2.2$
32	$T_{clk} \geqslant 14.4$	$4.8 \geqslant 2.2$
64	$T_{clk} \geqslant 15.7$	$5.7 \geqslant 2.2$
128	$T_{clk} \geqslant 17.7$	$6.6 \geqslant 2.2$

Taking into account all the previous data, the optimum architecture in terms of sensitivity should have $N_{\mu c} = 1$ and $M_{spl} = 32$. At this point, $N_{spad} \approx 390$, $A_{spad} = 3800\,\mu m^2$ and $A_{aspad} = 3100\,\mu m^2$. This configuration provides the best

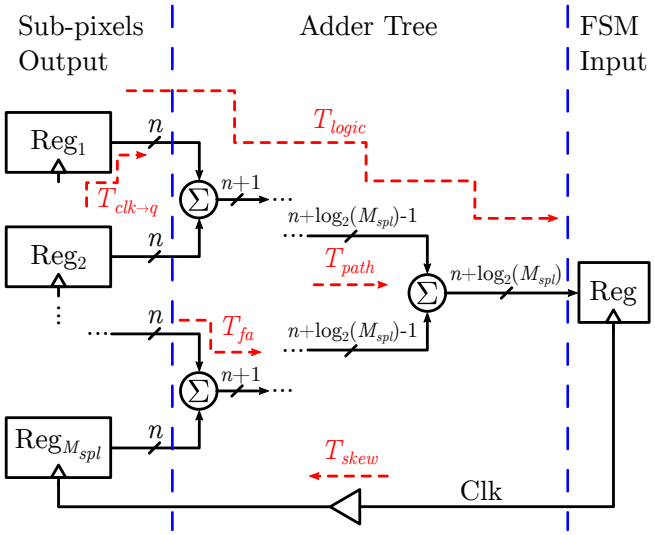


Figure 4.9: Data path from sub-pixel output to finite state machine (FSM) input.

sensitivity of all possible, however, its linearity is not optimal. Better linearity can be obtained with smaller SPAD, as it will be see in the next section. Moreover, from the yield point of view, $N_{\mu c} > 1$ are preferred to keep DCR as a linear function of active area and to reduce the impact of defective SPADs over the fill factor, that is, a small defective SPAD has smaller impact in the fill factor than a big one. As a consequence of the minimal variation of SNR as a function of M_{spl}, the SNR is not considered during the comparison.

4.2.2 Constant SPAD size

Another commonly used approach is to set a constant, and previously tested, SPAD size to find the optimum design point for a given FOM. This provides sub-optimum performance in terms of the FOM. However, the main advantage is that the whole performance can be better estimated if a complete characterization of the SPAD was carried out previously. This is because the size have an impact in the SPAD yield, e.g. bigger SPADs have worst jitter and DCR. Since the size is set constant, the analysis is reduced to study how the FOMs vary as a function of M_{spl} and $N_{\mu c}$. This simplifies the whole process, because no optimization is required and expressions (4.8) and (4.11) can be used to determine the sensitivity and SNR, respectively. The proposed SPAD dimensions for this analysis were presented in chapter 2: $x = 5\mu m$, $r = 5\mu m$, $GR = 1\mu m$, $A_{aspad} = 200\mu m^2$ and $A_{spad} = 449\mu m^2$. These dimensions are chosen to have a SPAD fill factor close to 40 % and keep the impact of noisy devices low.

As figure 4.10 represents, the optimum architecture in terms of sensitivity

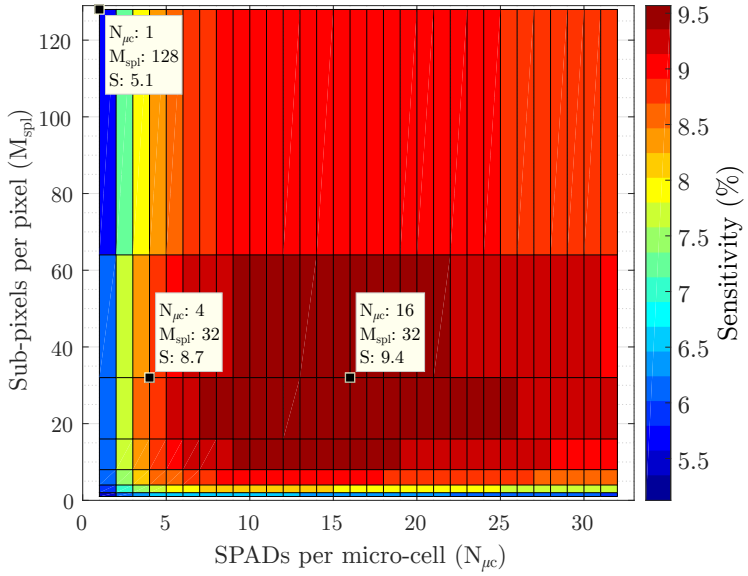


Figure 4.10: Sensitivity as a function of number of sub-pixels per pixel and number of SPADs per micro-cell for constant SPAD size.

is around $M_{spl} = \{16, 32\}$ and $N_{\mu c} = \{15, 24\}$. Contrary on what was shown in figure 4.4, there is a sub-optimum point if the SPAD size is constant. This is attributed to the fact that, for the given SPAD size, a slight fill factor increment have more impact in the final sensitivity than a slight decrement of the losses caused by sensor saturation. This is valid until saturation starts to rise for $N_{\mu c} > 24$.

As it was expected, figure 4.11 shows how a smaller SPAD size produces an increase in the sensor linearity at the price of lower sensitivity compared with the data of figure 4.5. Since the linearity is improved as figure 4.12 represents, the sensor could be useful for higher brightness applications than PET. This results illustrates the trade-offs between sensor linearity and sensitivity.

From figure 4.13, it can be observed that the SNR shows a similar behavior for both design approaches, variable and constant SPAD size. This is a consequence of a lower active area per pixel susceptible of producing dark counts at the points with the highest M_{spl} and lowest $N_{\mu c}$. If the DCR were independent of active area, the SNR would be proportional to the sensitivity, as expressions (4.8) and (4.11) show.

In this case, the hold-on and set-up conditions are summarized in table 4.3, these results are similar to those obtained in the previous section. Because both hold-on and set-up violations are determined by shortest and longest delay path, respectively, between the sub-pixel output registers and the FSM input registers. The only difference comes from number of bits of the sub-pixel

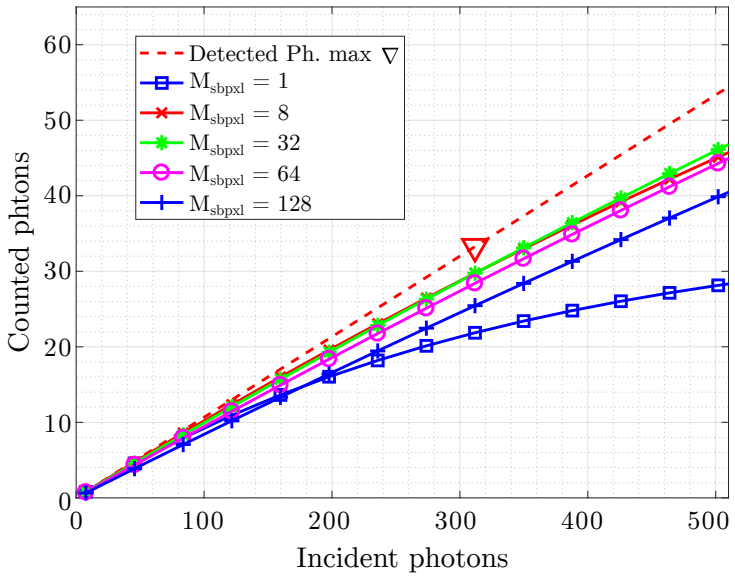


Figure 4.11: Counted photons as a functions of incident photons with 16 SPADs per micro-cell for constant SPAD size. The dashed line represents the detected photons (ideal sensor response), the triangle is pointing the maximum number of incident photons in T_{clk}.

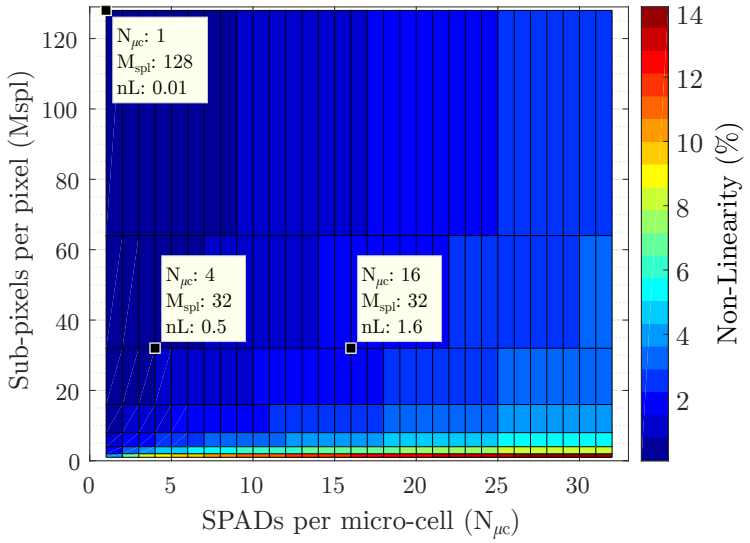


Figure 4.12: Pixel non-linearity for constant SPAD size.

counter, that is different

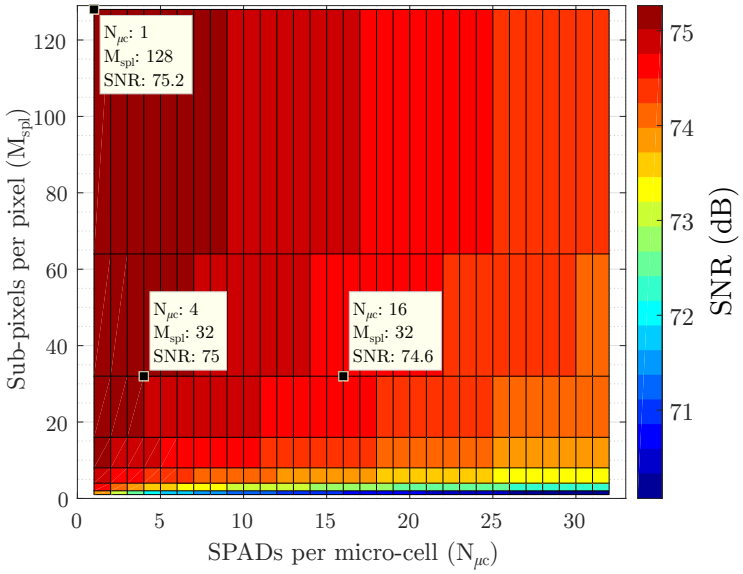


Figure 4.13: Signal to noise ratio as a function of number of sub-pixels per pixel and number of SPADs per micro-cell for constant SPAD size.

Table 4.3: Set-up and hold-on conditions for constant SPAD size and $N_{\mu c} = 16$.

M_{spl}	Set-up (ns)	Hold-on (ns)
1	$T_{clk} \geqslant 6.5$	$0.3 \not\geqslant 2.2$
2	$T_{clk} \geqslant 7.8$	$1.2 \not\geqslant 2.2$
4	$T_{clk} \geqslant 9.1$	$2.1 \not\geqslant 2.2$
8	$T_{clk} \geqslant 10.4$	$3.0 \geqslant 2.2$
16	$T_{clk} \geqslant 12.4$	$3.9 \geqslant 2.2$
32	$T_{clk} \geqslant 13.7$	$4.8 \geqslant 2.2$
64	$T_{clk} \geqslant 15.7$	$5.7 \geqslant 2.2$
128	$T_{clk} \geqslant 17.7$	$6.6 \geqslant 2.2$

Considering the sensitivity results of figure 4.10, the optimum architecture should have $N_{\mu c} = 16$, $M_{spl} = 32$ and $N_{spad} \approx 3160$. Again, the SNR is not taken into account during the comparison, because it is more or less constant, around 2 dB variation in the whole design space. This configuration gives better linearity than the proposed in the previous section, but its sensitivity is sub-optimal. The relatively low sensitivity is caused by the low fill factor of the SPAD itself because the SPAD guard-ring occupies a huge fraction of the SPAD area. Moreover, with $N_{\mu c} = 16$, the impact of a noisy SPADs can be attenuated by turning them off without having a noticeable loss on the array fill factor.

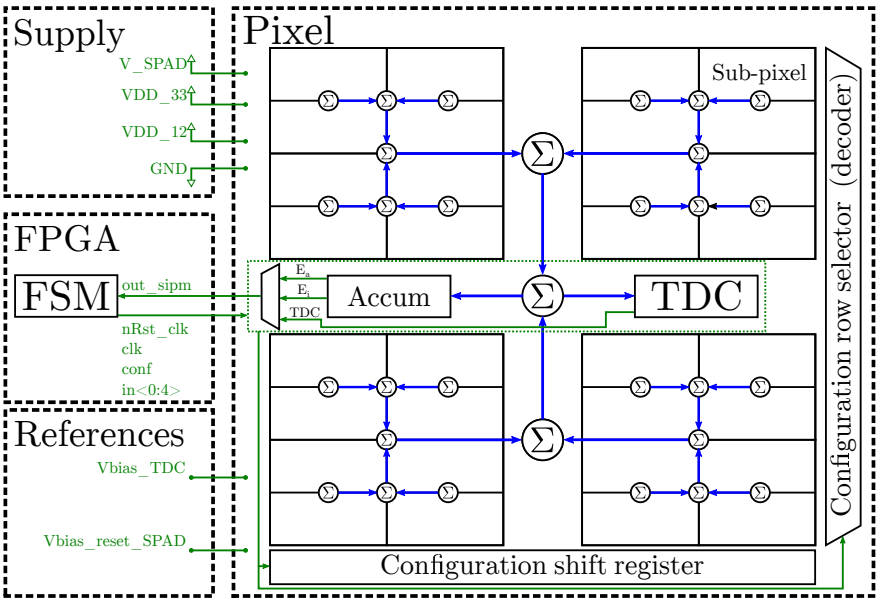


Figure 4.14: Proposed digital SiPM architecture with an adder tree of 5 levels and the SPADs are clustered into 32 sub-pixels.

4.3 Proposed Pixel Architecture

In this chapter, two architectural exploration processes have been proposed to study the main trade-off during the design of a digital SiPM. The principal variables studied here were the SPAD size, x; the number of sub-pixels per pixel, M_{spl}; and the number of SPADs per micro-cell, $N_{\mu c}$. Temporal constrains were also studied to verify the feasibility of these architectures in worst case corners, to guarantee no hold-on and set-up violations. The constant values used to solve the analytical expression were extracted from simulation of our library block, consequently, these results should not be taken as a closed results because there is room for further optimization at the block level, e.g. TDC conversion time, AQRC or SPAD active area.

Throughout the chapter, it has been shown that the number of sub-pixels per pixel have a noticeable effects on the sensitivity, because these blocks are responsible of losses, due to temporal compression. Variable and constant size optimization process have shown that a number of sub-pixels per pixel lower than 16 has an unacceptable value of sensitivity for a pixel size of 1.69 mm^2. On the other hand, if the number of sub-pixels per pixel is higher than 16, the losses due to temporal compression for a pixel size of 1.69 mm^2 are almost negligible. Furthermore, when the SPAD size is constant, and relatively small, large number of sub-pixels have a negative impact on the sensitivity, because of

a reduction of the fill factor. Additionally, the maximum number of sub-pixels is limited to 32 by set-up violations in both optimization process. For that reason, the proposed SiPM architecture should have 32 sub-pixels per pixel. figure 4.14 represents the top level diagram of the pixel. The pixels is composed of 32 sub-pixels; an adder tree of 5 levels; a TDC; an accumulator to integrate the number of photons; a multiplexer to choose between the TDC output, the number of detected photons during a clock period (E_i) an the accumulated number of photons (E_a); a shift register and a decoder to configure the SRAMs, i.e. to determine which SPADs are active during the sampling phase. Besides, the control and configuration of the pixels is implemented externally with a FPGA to have enough flexibility to test the first prototype of SiPM. Moreover, bias references for the TDC ring oscillator and SPAD dead time are externally controlled, note that a final implementation should have a internal PLL and band-gap references to compensate PVT variations as in [35, 36, 45].

The SPAD size for the proposed d-SiPM proof of concept have been selected to be $x = 5\mu m$ and $A_{aspad} = 200\mu m^2$. Although this means a non-optimum sensitivity, the decision is justified by two reasons, yield and jitter. On the one hand, the yield, in terms of excessive DCR and breakdown voltage uniformity, has not been studied for the SPAD sizes proposed by figure 4.3 in this technology. This may seem trivial, but a full characterization of large SPADs have to be carried out before implementing any SiPM, because excessive DCR or non uniform breakdown voltage could lead to an inoperative chip. On the other hand, as it was discussed in chapter 2, smaller SPADs have lower jitter than larger SPADs, since the avalanche propagation time is less sensitive to the position at which the electron-hole pair is created [32, 50, 52].

The optimization processes have proposed two very different number of SPADs per micro-cell as optimum values in terms of sensitivity. This disparity is attributed to the sensitivity trade off between fill factor and spatial compression. Accordingly to the constant size optimization, the number of SPAD per micro-cell should be 16 to obtain the largest sensitivity. However, such a large number has a negative impact on the jitter for two main reasons. Firstly, AQRC outputs connected to the same node increase the capacitive load and the signal rise time, leading to an increase in jitter as it was discussed in chapter 2. Secondly, it would not be trivial to equalize the signal path of 16 AQRC outputs connected to the same node. This is relevant to reduce jitter, because the arrival time of the trigger signal to the TDC must be independent on which SPAD has detected the photon. Consequently, the number of SPADs per micro-cell logic is set to 4, as a trade off between sensitivity, jitter and layout routing. Figure 4.15-(a) illustrates sub-pixel architecture and the the floorplan of the micro-cell and its connections with the counter block. With this configuration, the number of micro-cells per sub-pixel is 19 and the number of SPADs per sub-pixel is 76. Additionally, figure 4.15-(b) shows the block diagram of the counter block, the counters must be working interleaved to count all the monostable pulses, independently of the asynchronous arrival of photons. However, as in the case

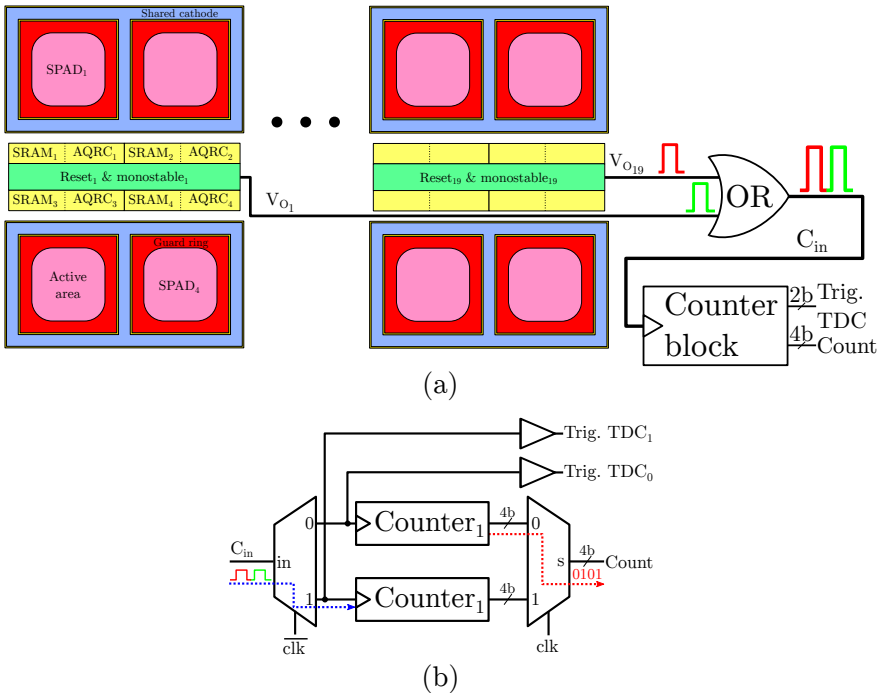


Figure 4.15: (a) Sub-pixel architecture, (b) block diagram of counter block working interleaved. Dotted arrows represent the data flow when clk = 0.

of the TDC, there is always a small misalignment of clock signals that leads to a dead time in the transition between sampling and reading.

The pixel micro-photography is shown in figure 4.16. The pixel core has an active area of 1.69 mm^2 and it is composed of 32 sub-pixels and the adder tree. The counter blocks and the adder tree are packaged into 8 distributed areas at the abutted corners of the sub-pixels. The output registers, the TDC and the configuration logic are allocated at the periphery of the core. Table 4.4 summarizes the architecture and the expected pixel performance, experimental characterization will be carried-out and discussed in the following chapter.

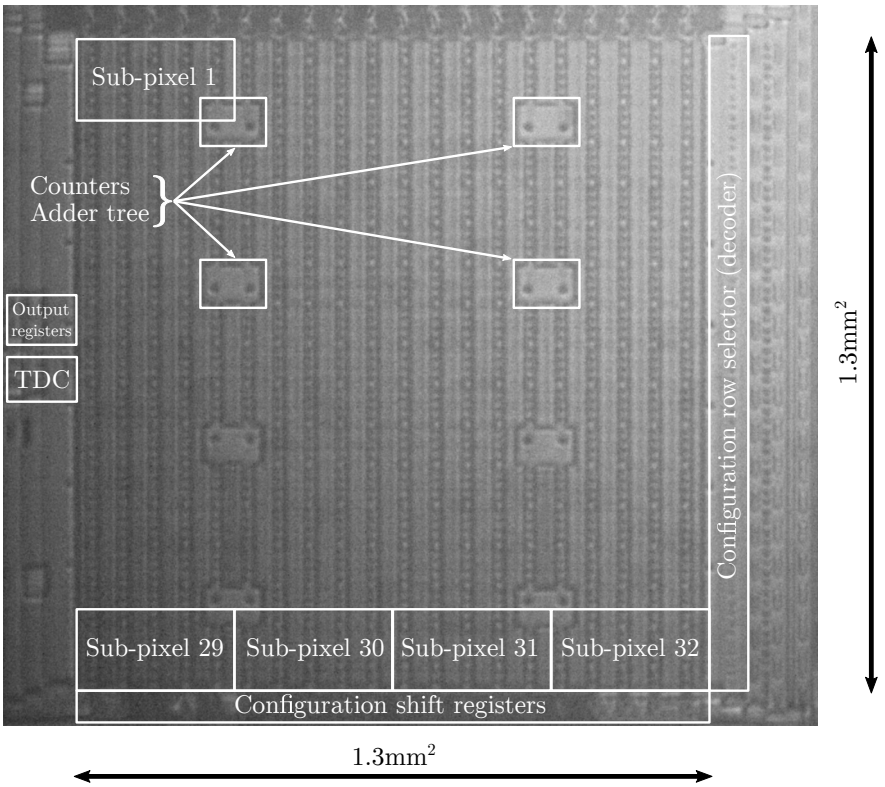


Figure 4.16: Pixel micro-photography.

Table 4.4: Expected performance of the proposed SiPM architecture.

Parameter	Value	Unit
A_{pxl}	1.69	mm^2
M_{spl}	32	-
$N_{\mu c}$	4	-
N_{spad}	2432	-
A_{spad}	450	µm^2
A_{aspad}	200	µm^2
FF	30	%
PDE(@420nm)	9	%
SNR	75	dB
Non-linearity	0.5	%

CHAPTER 5

DIGITAL SILICON PHOTOMULTIPLIER CHARACTERIZATION

As in many other fields of science, in microelectronics, it is common that the characterization of the first prototype is not only carried out to demonstrate its functionality but also to obtain the necessary information to improve future designs. If the initial assumptions were wrong, the root of the mistake is tracked and solutions are proposed, in order to verify them in future designs. Nowadays, the design of integrated circuits generally does not requires several error and trial iterations. Mainly thanks to the large panoply of computer-aided design (CAD) tools, the designers have at their fingertips and to the foundry efforts to provide well modeled, tested and verified devices. Nevertheless, this is not usually the case for SPADs in CMOS technologies, because foundries do not usually provide them as a standard library cells. The lack of a precise device model involves designing front-end on the basis of the acquired knowledge from the bibliography or the experience of previous implementations. Most of the times, this means to explore and to push the limits of the technology. Despite of the effort of the community to characterize these devices and show their functionality, the lack of experience in this field is an entry barrier for all those new designers that look with distrust, not without reason, any full-custom device that must be biased beyond the break down voltage of its p-n junction.

The main goal of this chapter is to characterize the implemented d-SiPM under low-light conditions, similar to the working environment in the application scenario. First of all, the I-V diode characterization is carried out in section 5.1 to extract the breakdown voltage and the internal breakdown resistance. In section 5.2, the array noise is obtained to determine the yield of the devices. Next, in section 5.3, the electro-optical characterization is carried out to find

the PDP. Finally, simulations results are provided to show the validity of the expressions obtained during the architectural exploration.

5.1 Diode Characterization

The first step of SPAD characterization is always to determine the diode breakdown voltage. Therefore, during the characterization of the other parameters, these can be easily expressed as a function of excess voltage. In order to verify the uniformity of the breakdown voltage, 400 devices from 4 different chips were studied. Low variability of breakdown voltage, both device-to-device and chip-to-chip, is required for SPAD structures implemented in arrays, to guarantee uniform biasing in the whole structure. Also, the shape of the I-V characteristic provides some qualitative information about the presence of PEB and the quality of the p-n junction. Firstly, the I-V characteristic with smooth slope, like the shown in figure 5.1-(a) for PP/NW3V3 devices, usually indicates the presence of PEB, because the breakdown resistance in these devices is higher than the breakdown resistance for a p-n junction without PEB. This higher resistance is caused by a smaller area (lateral junction) through which the current flows during the avalanche. Generally, lateral junctions are vertical p-n junctions at the perimeter of the active area, close to the guard ring. As it was explained in the chapter 2, devices with this characteristic are, in practice, insensitive to light, since the active area never reaches a proper breakdown voltage. Secondly, since all the SPADs share the same cathode, the variations in the breakdown voltage means a variation in the excess voltage for each device, leading to variations of PDP and DCR. figure 5.1-(b) represents the I-V characteristic of PW3V3/DNW diodes of different shapes and sizes. In this case, current through the device rise abruptly from 20 nA, due to reverse saturation current, to 4 mA. This last value is limited by the measurement instrument to avoid device destruction. Also, the current start to saturate around 100 μA because of the multiplexing circuitry of the test array.

The figure 5.2 illustrates the breakdown voltage of 400 PW3V3/DNW devices from 4 different chips. The breakdown voltage was defined as the voltage where the current increment has its maximum in the I-V characteristics. It can be observed that the breakdown voltage is around 18.0 V at room temperature. Moreover, in the worst case scenario, the variation of breakdown voltage is below 0.3 V all the SPADs. This entails that in an array of SPADs biased with an excess voltage of 3.0 V above the average, the excess voltage difference between the SPADs with highest and lowest breakdown voltage is below 10 %. And it is less than 5 % if the excess voltage is 6.0 V. The uniformity of the biasing helps to keep the uniformity of parameters like PDP and DCR, for this reason high excess voltages are preferred.

On the other hand, the figure 5.3 represents the breakdown voltage as a function of the chip temperature, controlled with an air forcing system. The

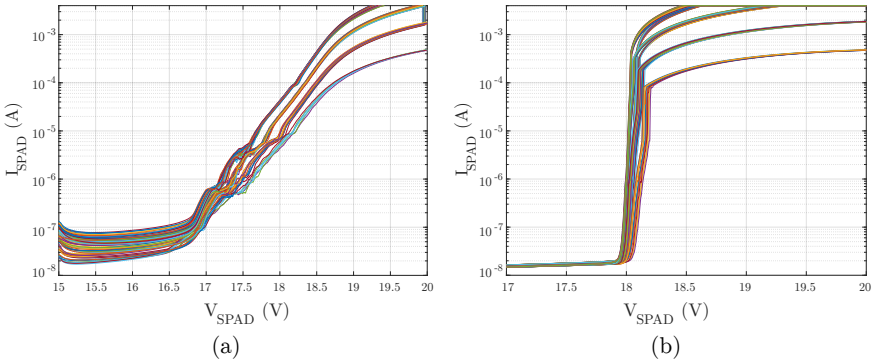


Figure 5.1: Inverse current as a function of cathode-anode voltage for SPADs with different sizes and shapes at room temperature: (a) PP/NW3V3 diodes with premature edge breakdown (PEB) and (b) PW3V3/DNW diodes without PEB.

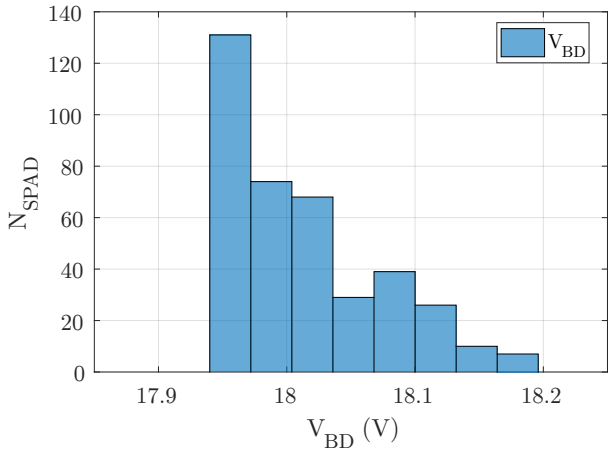


Figure 5.2: Histogram of break down voltage of the PW3V3/DNW devices at room temperature. The mean is 18.01 V and the total number of samples is $N = 400$ from 4 different chips.

temperature coefficient is around 14 mV/K for the range of 255–365 K, this value agrees with the reported results from [55, 56]. Qualitatively, this can be explained because charge carriers with higher temperature lose a fraction of their energy to optical phonons via scattering. These collisions cause a reduction of the ionization ratio of the charge carriers crossing the depletion region [33]. For this reason, higher electric field and, therefore, voltage is required to ensure the breakdown condition described in chapter 2. This data is necessary to emphasize the importance of a properly sized quenching resistor, since too large or improperly quenched avalanche currents may lead to SPADs self-heating due

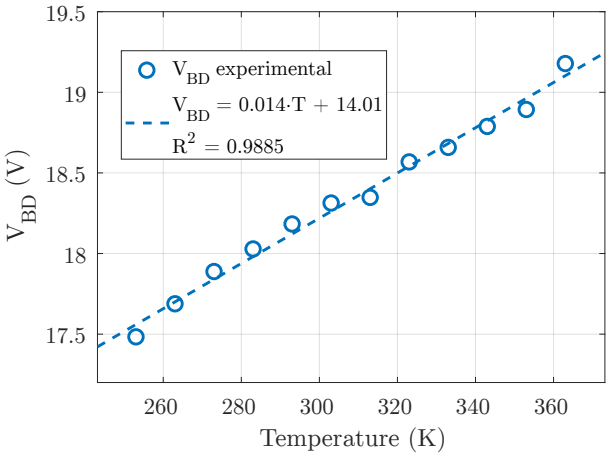


Figure 5.3: Break down voltage for a PW3V3/DNW SPAD as a function of temperature. Linear fit shows a slope of 14 mV/K.

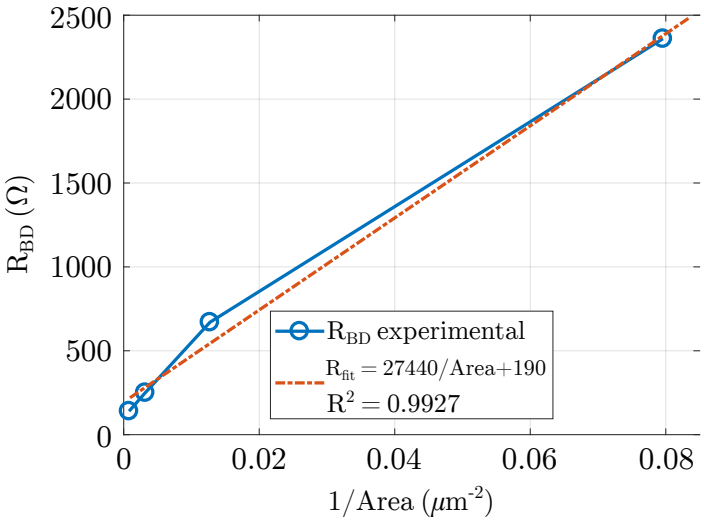


Figure 5.4: Breakdown resistance as a function of the inverse of the SPAD active area at room temperature.

to Joule effect. In a compact array, this means the creation of a hot spot able to reduce the excess voltage of its neighbor devices and, therefore, to create variations in the PDP and DCR of the array.

Also from the I-V characteristics, the breakdown resistance (R_{BD}) can be calculated. Despite only 4 different SPAD sizes were implemented, this data will help to improve the electrical models for future designs. In figure 5.4,

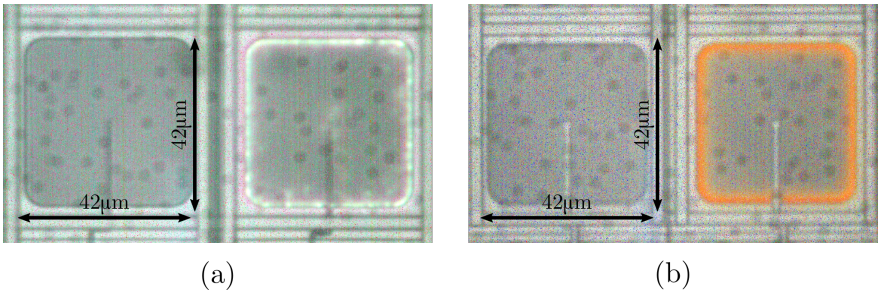


Figure 5.5: (a) Avalanche photo-emission of PP/NW1V2 SPAD with premature edge breakdown and I_{SPAD} = 10 mA (b) avalanche photo-emission of PW3V3/DNW SPAD without premature edge breakdown and I_{SPAD} = 19 mA. Turned-off SPADs are shown for comparison purposes.

the value of the slope (~30 kΩ/µm^2) is high, but it is consistent with the fact that Deep N-well layers have a low doping. The offset of 190 Ω is due to the parasitic resistance of the physical set-up and the multiplexing logic. As it was shown in chapter 2, the time constant of the quenching can be approximated to a linear function of the breakdown resistance and the junction capacitance ($\tau_Q \approx R_{BD}C_{KA}$). Apart from increasing the quenching time, less doped layers are also associated to devices with larger jitter than devices implemented with more doped layers (less resistive junctions), like the PP/NW3V3. This larger jitter is the result of statistical variation in the number of emitted photons during the avalanche building-up and spreading [52].

A common experiments to test the uniformity of the p-n junction consist of studying the avalanche photoemission, that is, the emission of photons in the visible spectrum due to the impact ionization of electron and holes in the p-n junction [47, 51, 61]. This technique is very useful to identify regions where avalanches are occurring, because these regions are highly illuminated. This helps to study the electric field uniformity across the active area, the effectiveness of guard ring structures or the presence of PEB [47, 123, 124, 125]. To bias the diodes in a self-sustainable avalanche region, low quenching resistors are required, this leads to reverse currents in the order of 1–20 mA. In figure 5.5-(a), photoemission attributed to PEB and non-uniform doping can be observed at the borders and the center of a PP/NW1V2 diode, respectively. This image was obtained by superimposing an black and white image of the diode with background light over a color image of the same region without any illumination.s Long exposure, around 14 seconds, was required to obtain the picture of photoemission. The current through the device is 10 mA. The reason why there is discontinuous glow at the border is because of the lack of electric field uniformity in the PP/NW1V2 depletion region. To avoid the problems arising from superficial highly doped layers, like the previous ones, designers tend to choose low-doped layers, since they have a more uniform doping profile [48]. In

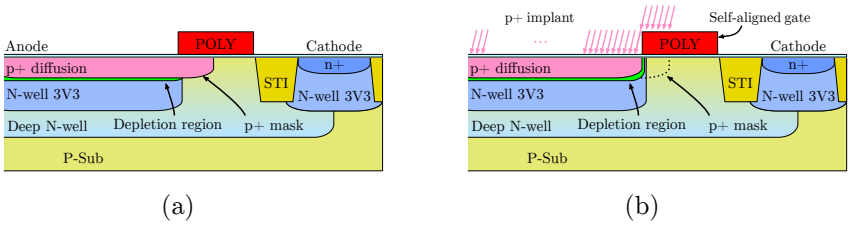


Figure 5.6: (a) Desired PP/NW3V3 SPAD structure (b) implemented PP/NW3V3 SPAD structure.

the case of PW3V3/DNW diodes, photoemission is much more uniform along the borders as figure 5.5-(b) illustrates, even for currents as high as 19 mA. Under such bias, a very slight glow can be appreciated at the center of the device. This suggest that a large fraction of the avalanche current is flowing through the guard ring, instead of through the bottom of the device.

Qualitatively, these results along with the I-V characteristics indicate that the implemented PP/NW devices are not proper candidates to be part of an array, because of the low electric field uniformity in the p-n junction and the presence of PEB. Nevertheless, devices with equivalent layers are reported in [56], showing promising results in terms of PDP, DCR, and jitter. As it was discussed in section 2.3, an error during the design of PP/NW3V3 structure leads to PEB. The p+ diffusion mask was extended to enclose the N-well as figure 5.6-(a) illustrates. However, this enclosure was done below the poly-silicon, where the poly-silicon acts as a stopping layer, preventing the formation of the diffusion below it as figure 5.6-(b) shows. To overcome this issue in future designs, the poly guard-ring should be reduced or moved away to allow the proper formation of p+ diffusion. On the other hand, the photoemission in the PW3V3/DNW devices seems to be uniform at the borders, but, since low glow is appreciated a the center, a more effective guard ring should be implemented to guarantee that the avalanche is well confined in the active area and no avalanches occurs at the borders.

5.2 Dark Count Rate

The DCR characterization of the SiPM provides information about the spatial distribution of noisy diodes and its yield. For obvious reason, this characterization must be carried out SPAD by SPAD under complete darkness with different excess voltages to study the variation of this parameter. Furthermore, the dead time of the devices is set to 5 µs, a value high enough to reduce the effect of after-pulsing. Although the magnitude of the dead time is small, the measured DCR must be corrected to estimate the real value of DCR [126]:

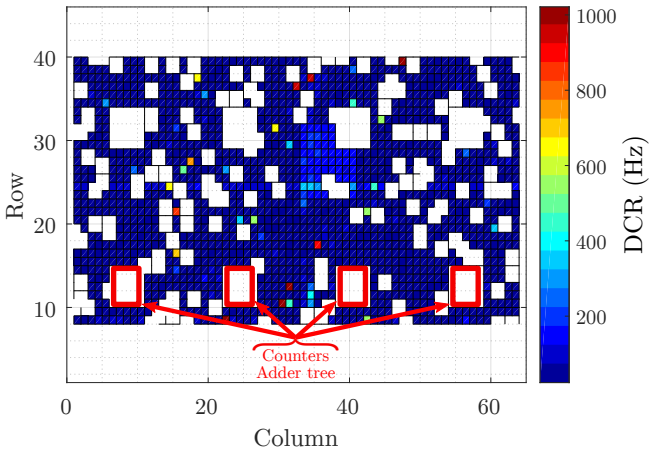


Figure 5.7: Dark count rate (DCR) map for PW3V3/DNW devices with $V_{ex} = 2\,\text{V}$ and $t_d = 5\,\mu\text{s}$. White areas stand for non-working quenching ($I_{SPAD} > 1\,\text{mA}$).

$$DCR_r = \frac{DCR_m}{1 - DCR_m\ t_d} \tag{5.1}$$

where DCR_r is the estimated real value of DCR, DCR_m is the measured DCR and t_d is the dead time. Since the number of counts is small, the integration time of each sample is 12 seconds. Figure 5.7 represents the DCR for a SiPM with an excess voltage of 2.0 V. It can be observed that the DCR is below 200 Hz in almost all the SiPM, except for few devices that reach values over 1000 Hz. The areas with null data represents the SPADs without proper quench. As a result, these devices are always emitting photons and no counts are registered.

At the same time, figure 5.8 represents the DCR for a SiPM with an excess voltage of 3.0 V. As it was expected, the current through the SPAD increases at higher excess voltages, consequently, a larger number of quenching circuits are unable to quench the avalanche. Under such bias conditions, the yield degradation can be attributed mainly to the bad design approach used to size the quenching transistor of the AQRC. The design approach followed entails that the avalanche is only quenched after the quenching transistor is switched off, because conductive quenching transistors are used to better detect the avalanche current peaks [77]. This sets a large avalanche current that modifies the temperature and parameters of the SPAD, the quenching transistor and the comparator. As a result, if the actual excess voltage of the SPADs is reduced, the comparator, in this case implemented with a wide transistor, can not detect the avalanche and switch off the quenching transistor. This leads to a self-sustainable avalanche.

While it is true that these DCR maps are indispensable to identify faulty devices that must be turn-off during the SiPM normal operation, it is also true

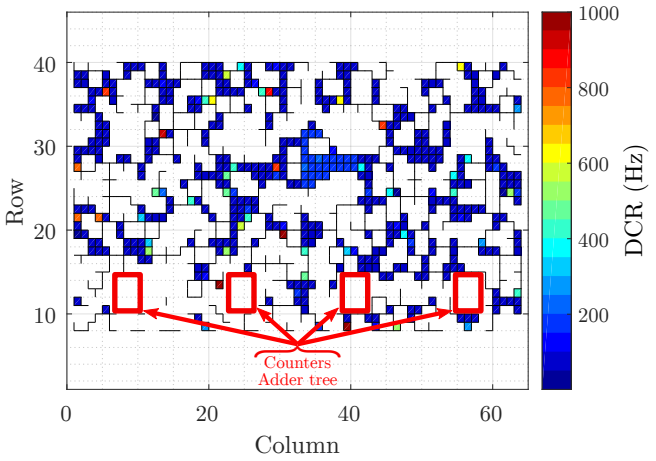


Figure 5.8: DCR map for PW3V3/DNW devices with $V_{ex} = 3\,\text{V}$ and $t_d = 5\,\mu\text{s}$. White areas stand for non-working quenching ($I_{SPAD} > 1\,\text{mA}$).

that it is complicated to study the yield quantitatively. For this reason, DCR is also represented as a cumulative histogram of population. This kind of plots are very useful to interpret which percentage of the total number of SPADs has a DCR equal to or below to a given value. The figure 5.9 shows the cumulative DCR histogram for all the SPADs of a SiPM for different excess voltages and dead time of 5 µs, to avoid after-pulsing. The median DCR for the devices biased at 3 V is 80 Hz, which means a median of 0.4 Hz/µm^2. This value is higher than the reported in the state of the art d-SiPM implemented in CMOS technologies [56, 55]. With an excess voltage of 2 V, the percentage of faulty devices is around 10 %, while for excess voltages of 3 V, this percentage rises up to 35 %. This low yield can be attributed to device self-heating caused by the high currents through the SPADs during avalanche. This, at the same time, means that the number of defective SPADs increases for long integration times, since the self-heating drifts the micro-cell (SPAD and quenching circuit) characteristics over time.

In order to validate the previous hypothesis, it is worth studying the effect of temperature on the DCR. To compensate the excess voltage variation with temperature, the data obtained in the previous section was used. The experimental data from figure 5.10 can be fitted to the Arrhenius expression to estimate the activation energy (E_A) of the charge carriers responsible from triggering the dark counts:

$$DCR = A \exp\left(-\frac{E_A}{k_B T}\right) \tag{5.2}$$

where A is a constant, E_A is the activation energy, k_B is the Boltzmann constant and T is the absolute temperature in Kelvin. It can be see that for temperatures

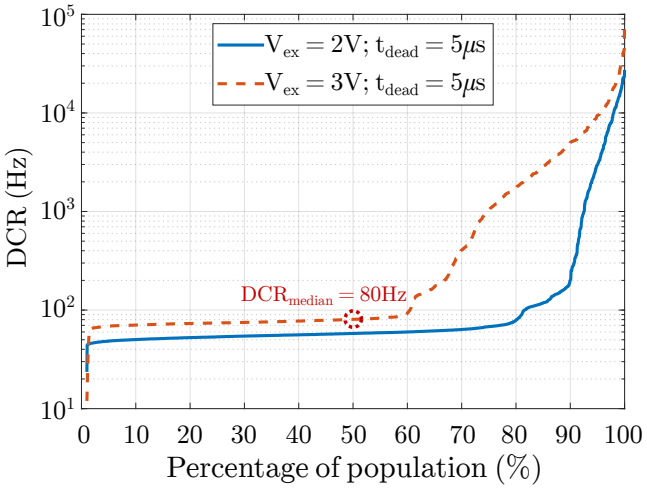


Figure 5.9: Cumulative DCR population for 2432 PW3V3/DNW SPADs at room temperature. SPAD active area is 200 μm^2 and median DCR 0.4 $Hz/\mu m^2$.

higher than room temperature, the activation energy is similar to silicon band-gap (1.12 eV). Hence, thermal trap assisted Shockley-Read-Hall carriers are responsible of the largest component of DCR [127]. Moreover, it can be observed that the DCR multiplies it value by ten every 25–30 K over room temperature. It seems feasible that an increment of p-n junction temperature due to poorly quenched avalanches results in the DCR increment that can be observed in figure 5.9 between 80–90 % of the the population. On the contrary, when the temperature is below room temperature, the activation energy is smaller and the main component of DCR is band-to-band tunneling [57].

Also, the effects of faulty quenching can be observed with the naked eye because of the photoemission and high current consumption. The figure 5.11-(a) shows a microphotography of a SPAD micro-cell in color. The figure 5.11-(b) shows the same SPAD micro-cell in black and white with a picture of photoemission superimposed. Long exposure, around 14 seconds, was required to obtain the picture of photoemission. The microphotography shows the light produced during continuous avalanches in two SPADs biased with and excess voltage of 1 V and a current of 1 mA. Such currents are incompatible with SiPM nominal operation, because the number of SPADs per pixel is 2432. Even with only 10 % of active devices, it is experimentally observed that the current consumption is so high that the bias can not be properly supplied because of voltage drop across the entire SiPM. As explained before, this implies that the comparator is unable to detect the avalanche current and to quench it. Once again, this points out that the approach followed to size the quenching transistor is not appropriate for a normal SPAD operation.

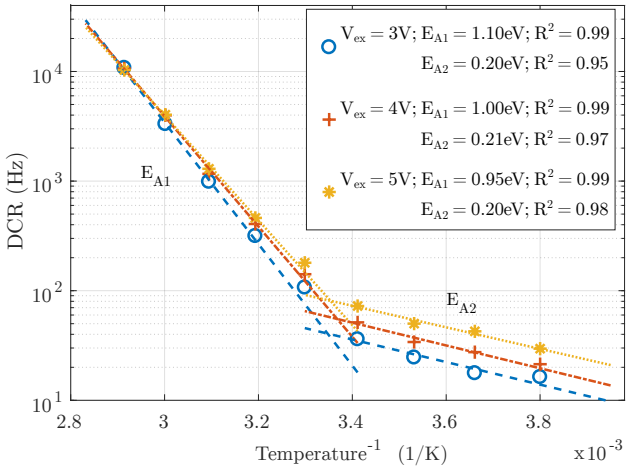
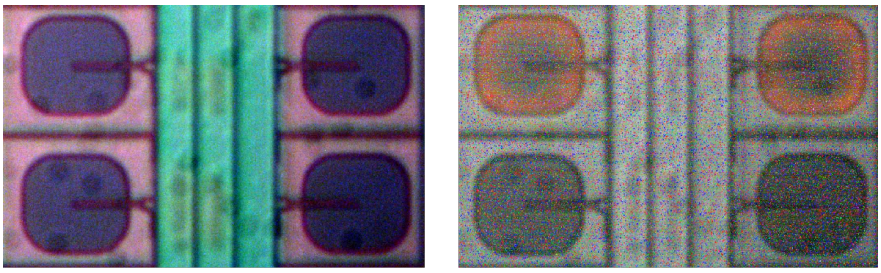


Figure 5.10: DCR as a function of temperature for PW3V3/DNW SPAD. E_{A1} slope is associated to thermal trap assisted Shockley-Read-Hall generation and E_{A2} slope is associated to band-to-band tunneling.



(a) (b)

Figure 5.11: (a) Color micro photography of a SPAD micro-cell and (b) avalanche photo-emission of two SPADs with out proper quenching ($V_{ex} = 1\,V$ and $I_{SPAD} = 1\,mA$), the other SPADs are turned-off.

5.3 Photon Detection Probability

The ability of the sensor to detect and count the incident photons provides valuable information about the performance of the technology and the proposed architecture. In this case, it is not possible to characterize the SiPM as a such, since only a small number of SPADs have a stable behavior over time. Besides, this reduces considerably the number of experiments that can be performed to obtain significant, representative or relevant information about the SiPM. Parameters like PDP can be obtained, however, the after-pulsing and cross-talk

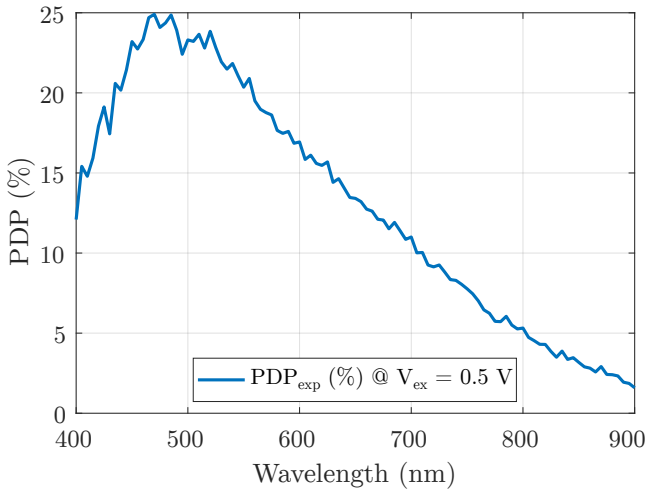


Figure 5.12: Photon detection probability as a function of wavelength with PW3V3/DNW SPADs for $V_{ex} = 0.5\,\text{V}$ and $t_d = 5\,\mu\text{s}$.

loose their validity in a design with such problems.

The PDP measurement was carried out with an Ebert-Fastie monochromator under single photon regime $\left(\sim 1\,\text{nW/mm}^2\right)$ to avoid the effects of pile-up in the sensor. Otherwise, multiple photons could impact the sensor at the same time (or during its dead time), keeping the ratio between detected and incident photons lower of what it should be. The PDP can be calculated as:

$$PDP = \frac{C_{out} - DCR}{OP_{in}} \tag{5.3}$$

where C_{out} is the number counted avalanches at the SPADs, DCR is the SPAD dark count rate and OP_{in} is the number of incident optical photons in the SPAD obtained from a calibrated photodetector. The counts of individual SPADs were recorded and its valued averaged. In figure 5.12, the PDP is illustrated as a function of the wavelength at an excess voltage of 0.5 V. The PDP peak is 25 % and can be found at 470 nm, this value is relatively high taking into account the low excess voltage and the kind of p-n junction. In the near infrared, the PDP is still high, in the order of 2.5 % at 900 nm. No further measurements at higher excess voltages could be performed due to device malfunctioning.

By selecting the devices that can be biased at higher excess voltage without getting stuck at avalanche, it was possible to take measurements of the PDP during short periods of time at different excess voltages as figure 5.13 illustrates. As it was expected, the PDP follows an exponential trend. The exponential fit of the PDP at its peak, $\lambda = 470\,\text{nm}$, reaches a value close to 45 % at 2.5 V excess voltage. On the other hand, for the $\lambda = 420\,\text{nm}$, the exponential fit saturates for values close to 2.5 V at 30 %. This last wavelength is relevant because the

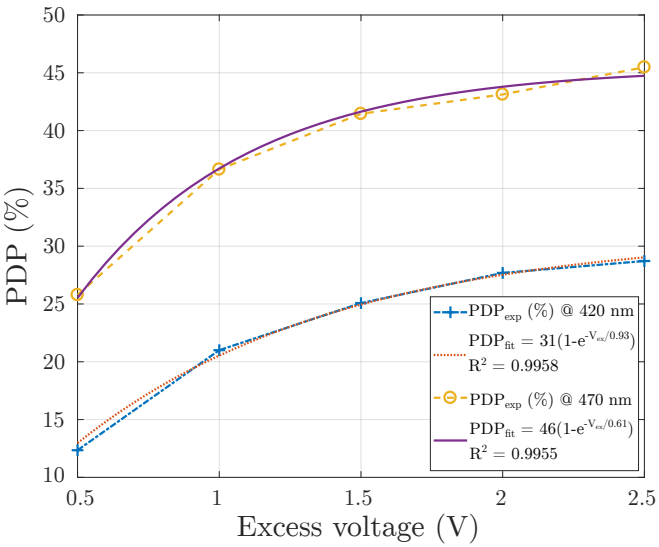


Figure 5.13: Photon detection probability as a function of excess voltage for PW3V3/DNW SPADs at $\lambda = 420\,\text{nm}$ (circle) and $\lambda = 470\,\text{nm}$ (cross).

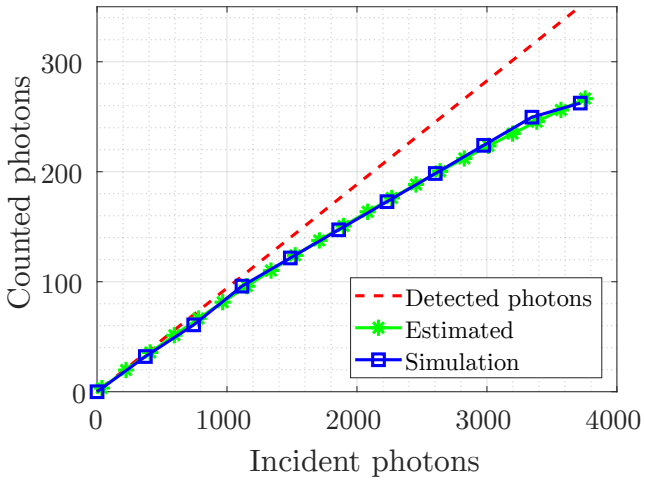


Figure 5.14: Counted photons as a functions of incident photons for the implemented pixel at $\lambda = 420\,\text{nm}$ and $V_{ex} = 3.0\,\text{V}$. The dashed line represents the detected photons (ideal sensor response), star marker represents the estimated detected photons and square marker represents the simulated detected photons.

emission peak of LSO and LYSO scintillators is centered at this wavelength.

Since it is impossible to activate a significant portion of the SiPM to verify experimentally the functionality of the proposed architecture, it has been decided

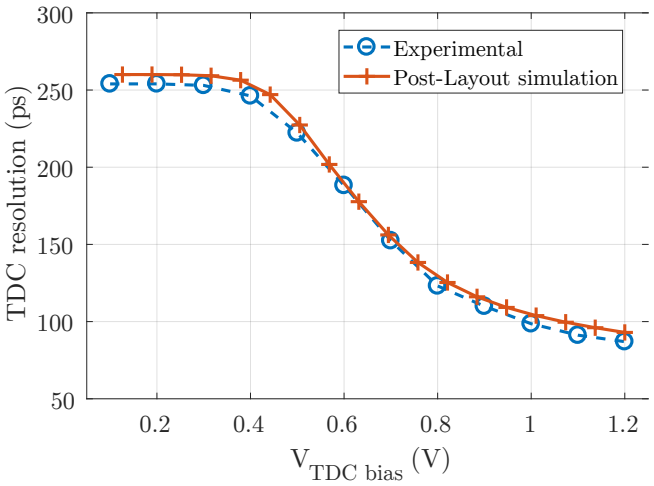


Figure 5.15: TDC resolution as a function of bias voltage. Solid lines represents post-layout simulation results and dashed line stand for experimental data.

to simulate the part of the circuit in charge of counting the photons. To emulate the random nature of the photon arrival, pseudo-random number generators were used at each simulation to determine the number of detected photons, the SPADs at which the photons are absorbed and their arrival time. The simulations were repeated iteratively until the error between the mean simulated value and the estimated value is less than 5 %. The figure 5.14 illustrates the ideal response of a sensor only taking into account the PDE with dashed line, the estimated value of counted photons from the expressions derived in chapter 4 with star markers and results obtained from simulations with square markers. It can be appreciated that the estimated value from expressions and simulations lie over the same trajectory. These results suggest that the expressions used during the architectural exploration are valid, despite this could not be verified experimentally.

5.4 Time to Digital Characterization

The TDC performance is determinant to obtain the arrival time of visible photons and, hence, to calculate the ToF of the gamma photons in PET. The characterization of this circuit is mainly carried out with a code density test. Thus, the non-correlated DCR from few active SPADs can be used as a trigger for the TDC. The main advantage of this technique is that no input signal is required and few pad test are saved. However, taking into account the current circumstances, this decision has reduce the test flexibility considerably. Some parameters, like resolution, single shot precision and jitter, are difficult or impossible to measure,

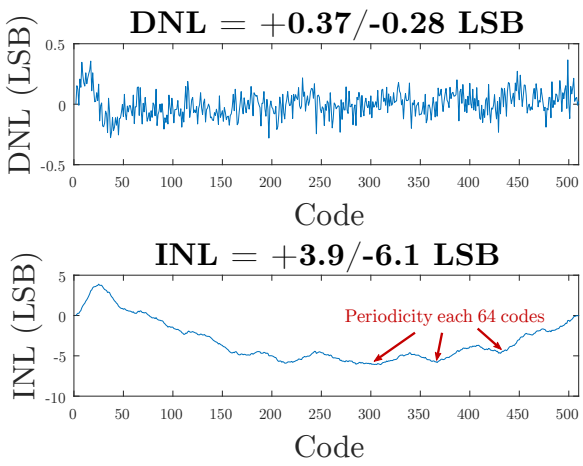


Figure 5.16: Differential non-linearity and integral non-linearity of the TDC of the first 512 codes (~ 50 ns). The periodicity at the integral non-linearity could be attributed to coupling between the ring oscillator and the ripple counter.

because the control signals cannot be introduced in this first prototype.

In figure 5.15, the time resolution of the TDC as a function of the bias voltage is represented. It can be observed that the differences between the post-layout simulation results and the experimental data are below 5 %. A time resolution has a tuning range from 250 to 90 ps, with a maximum power consumption of 1.1 mW. The right side of this characteristic shows a slope of ~190 ps/V.

The top of figure 5.16 illustrates the differential non-linearity (DNL) as a function of the output code for one of the TDCs implemented in a pixel when the time resolution of the converter is ~ 100 ps. The maximum DNL is around ±0.37 LSB in a range of 50 ns. The integral non-linearity (INL) is below ±6.10 LSB for the same range. Furthermore, the INL has a periodic behavior with a length of 64 codes, this may suggest the coupling between the ripple counter and the ring oscillator through supply and substrate. This highlights how critic is to isolate and put different supplies between the ring oscillator and the rest of the logic to reduce non-linearities. Future designs must take this into account and sacrifice some area to separate the supplies. The ring oscillator can be isolated from the rest of the circuit with deep N-wells.

5.5 Characterization Summay

Along this chapter, PP/NW3V3 and PW3V3/DNW SPADs implemented in 110 nm CIS technology were characterized. I-V and photoemission characterization have shown that PP/NW3V3 SPADs suffer from PEB caused by an undesired

poly guard-ring acting as an stopping layer of p+ implant. The PW3V3/DNW SPADs characterization shows that the breakdown voltage is around 18 V with a small variability, ±0.3 V for 400 devices tested. Furthermore, a good overall performance in terms of PDP ~ 46 % and DCR 0.4 Hz/μm^2 for an excess voltage of 3 V.

The main issue with the implemented SiPM is related with the inability of the AQRC to quench the avalanche due to its low quenching resistance. This leads to a large steady state current, in the order of mA, flowing through the devices. Such amount of current is unfeasible for large arrays, because of the voltage drops in the SPADs supply and the excessive device heating.

The time resolution of the TDC is below 100 ps and it agrees with the post-layout simulations in the whole TDC bias range. The TDC DNL is around ±0.37 LSB, while the INL is around ±6.10 LSB. Although the INL results are not as good as the DNL results, they can be corrected trough post-processing.

CHAPTER 6

RESULTS AND DISCUSSION

This chapter performs a critical assessment of the achievements of this thesis and their connections to the initial global goal. This goal was to design and characterize a functional proof-of-concept SiPM. This global goal called for a detailed architectural study, on the one hand, and for the design of the mixed-signal and logic basic building blocks required for the targeted function, on the other hand. Technology considerations constrained the quest for this goal. First of all, while 3D-stacked, hybrid technologies are a priori better suited for sensory-processing chips, their limited availability and robustness forced using planar technologies. Second, while standard very large-scale integration (VLSI) technologies are advantageous in terms of logic operator density, power and speed, they suffer from having extremely poor SPADs, with DCRs ~ 50 kHz. These considerations motivated the selection of an 110 nm CIS technology which, as shown in this thesis, enables SPADs with ~ 50 Hz DCR. Still, the characterization data available from the foundry are not complete regarding SPAD behavior. Also, the foundry does not provide technology parameters (such as doping profiles) to perform thorough device engineering using TCAD tools. Some design hypothesis remained hence un-certain until confirmation by measurement form working silicon. Unfortunately, such experimental characterization highlighted a flaw in the theory sustaining AQRC strategies. Basically, resistances were incorrectly estimated based on the foundry data, and this resulted in an incorrect control of positive feedback avalanche phenomena. The re-spinning of the SiPM is hence needed to correct system integration. However, both the architectural studies and the building blocks remain valid assets for such a re-spinning.

6.1 Main achievements

Architectural explorations of digital SiPMs have been carried out analytically. To do so, the key functions that a digital SiPM for ToF-PET must perform were identified, namely: to detect and count photons, to measure their arrival time, and to process the scintillation event. Then, optimum points in the design space were encountered. This search has led to the design of a SiPM for ToF-PET in a CIS technology. The complete functionality of the prototype could not be demonstrated under laboratory conditions, due to several flaws in the design of the AQRC. Still, the underlying problems are identified and the thesis contributions have been demonstrated valid. These contributions can be grouped into three different blocks: SPADs, AQRC, and SiPM architectural exploration.

6.1.1 Single Photon Avalanche Diode

Two SPAD structures have been implemented in an 110 nm CIS technology, whose depletion regions were composed of PP/NW3V3 and PW3V3/DNW layers. The SPADs were designed as parametric cells to simplify the sizing of the active area and the guard ring thickness. This has allowed the implementation of SPADs with different sizes and shapes to study the impact of these parameters in the performance of the device [128].

The characterization of PP/NW3V3 SPADs points out the presence of PEB, as was discussed in chapter 2. On the contrary, the SiPM implemented with PW3V3/DNW SPADs does not suffer from PEB. Characterization shows that the median DCR is around 0.4 Hz/μm^2 with an excess voltage of 3.0 V. Also, the PDP is estimated to be 45 % at 470 nm with an excess voltage of 3.0 V. These are promising results for a wide range of applications, like PET, LiDAR or FLIM to name few.

6.1.2 Active Quenching and Recharge Circuit

Despite the promising performance of PW3V3/DNW structure, the quenching circuits implemented in the SiPM are not working properly because of the design approach chosen leads to large avalanche currents, in the order of 1 mA, even with excess voltages below 1 V. This does not only affect the power consumption, cross-talk, photoemission and after-pulsing, but also the functionality of the digital SiPM itself. For this reason, it is impossible to activate a large number of SPADs at the same time to verify the performance.

This flaw is a consequence of the incorrect estimation of the currents and the operation principle of the AQRC circuit. This latter relies on the transition from saturation to cut-off of the quenching transistor used to stop the avalanche. Therefore, if the quenching transistor does not reach the cut-off state, the current through the SPAD can be orders of magnitude higher than the quenching current. Under these circumstances, avalanche photoemission is sustained over

time. The most straightforward solution is to size a quenching transistor longer than wider. This involves using a transistor with high output impedance in the saturation state to guarantee that the current through the SPAD is always below the quenching current. Similar sizing strategies can be found in [70].

6.1.3 Architectural Exploration

The SiPM architectural exploration has helped to identify the basic functional blocks of a SiPM for a ToF-PET scanner. These blocks support the extraction of information from a scintillation event to enable the reconstructions of the LOR and the annihilation point of the positron. On the one hand, the incident photons must be counted to infer the energy transferred by the gamma photon during its interaction with the crystal. On the other hand, the arrival time of these photons allows to estimate the ToF of gamma photons and, hence, their origin inside the body. Sensitivity is chosen as an optimization figure of merit for architectural exploration. The variables of the optimization are the SPAD size, number of sub-pixels, and the number of SPADs per micro-cell. With the results of the optimization, it is easier for the designer to justify architectural decisions.

To the best of the author's knowledge, the state-of-the-art does not include a similar technique to study the design space of digital SiPM and optimize its performance. However, this technique has some limitations. On the one hand, optimization process relies on the selected figure of merit and the design variables. If the figure of merit does neither have a maximum nor a minimum for the given variables, the optimization is not possible. On the other hand, this approach will be eventually overcome by cost-effective 3D-stacked technologies. In such a scenario, a SPAD custom technology might be used at the top tier, and deep sub-micron technologies might be employed for the front-end and digital processing at the second tier.

6.2 Beyond the PhD

6.2.1 Single Photon Avalanche Diode structure

Even though PW3V3/DNW SPADs have a good performance, there is still room for further improvements. Points calling for further analysis include the impact of the guard ring size on the DCR, after-pulsing, jitter, cross-talk, and PEB. Also, the benefits and drawbacks of the biased guard ring structures described in [66, 129] should be studied.

On the other hand, the PP/NW3V3 SPADs implemented suffer from PEB which calls for a redesign of the structure. By mistake, the poly layer was extended up to cover the p+ diffusion, which makes the edge of the depletion region to be inside the N-well, leading to PEB. The poly gate acting as a mask for

drain-source implants is a common feature call self-aligned gate in silicon-gate technologies. As a consequence, it was impossible to characterize the SiPM implemented with these devices. However, once the origin of the problem is identified, there are still some open questions about the performance of this structure for SiPM. A priori, this structure must have better PDP at shorter wavelengths, and smaller jitter than PW3V3/DNW SPADs, as [56] reports.

One of the aspects where more efforts are being made in the design of SiPM for ToF-PET is in the reduction of CTR. In this sense, the contribution of SPAD to the overall CTR is not negligible, in the order of 80–200 ps [128]. This means that there is still a lot of work at the circuit level to achieve the time accuracy that satisfies the needs of PET community [19].

6.2.2 Quenching and Reset Circuit

The quenching and reset circuit must have a fundamental role in the re-spinning. The most promising topologies in terms of power consumption and occupied area are the PQARC, since they control the maximum avalanche current with a current mirror. On the contrary, an AQRC needs larger quenching transistors than PQARC to achieve the same avalanche current. Also, these topologies allow higher excess voltage if cascode transistors are used. This means an improvement in PDP and jitter, at the cost of increasing the DCR.

6.2.3 Time to Digital Converter

As for TDC, although the implemented topology has a time resolution smaller than 100 ps, the non-idealities found during characterization provide clues for enhancements. The occupied area was reduced to keep the fill factor as high as possible at the expense of increasing substrate and supply noise coupling. The straightforward solution requires properly de-couple the supplies of the ring oscillator and the rest of the logic. Also, the substrate of the ring oscillator can be isolated from the rest of logic with deep n-wells.

For the application to benefit from the improvements in ToF, more accurate and precise circuits are required, even more than the implemented topology allows. The industry is pushing towards smaller time resolutions, in the order of tens of picoseconds. This seems to lead to sub-delay line architectures, however, the improvements in time resolution often comes at the expense of long conversion time, high power consumption and occupied area. One of the more attractive alternatives for future designs could be the local passive interpolation [87]. Nevertheless, power consumption and occupied area must be carefully evaluated.

6.2.4 New Silicon Photomultiplier

Results of the SiPM characterization have provided a lot of useful information for future designs. The technology seems to have potential to implement future designs based on SPADs, since it performs well in PDP and DCR. Owing to the low DCR, bigger SPADs can be implemented, like the proposed by the optimization process, to increase the fill factor and sensitivity. Experimental characterizations also highlight the shortcoming of the proposed design in terms of testability. In order to reduce these problems, control and test structures must be added in future design. This, along with the previous improvement, will lead to a functional SiPM for ToF-PET.

CHAPTER 7

CONCLUSIONS

This thesis represents a contribution to the design of single-photon detectors for medical applications. The detection of scintillation events, such as those produced inside of a LSO crystal in a ToF-PET scanner, raises two fundamental challenges. On the one hand, sensors must be sensitive enough to work in the single-photon regime. On the other hand, they must measure the arrival time of these photons with resolution in the picosecond range. In this regard, a SiPM architecture has been explored to study analytically the design variables that determine the performance of the sensor. This dissertation addresses different problems related to this architecture and makes the contributions summarized in the bullet points listed below.

- Two SPAD structures are selected and implemented to study their performance and assess their suitability for ToF-PET. The PP/NW3V3 structure was expected to have smaller jitter and better PDP for short wavelengths, thus providing better matching to the scintillator spectrum. However, the extension of p+ diffusion beneath the polysilicon guard ring leads to unwanted PEB. As for the PW3V3/DNW structure, the characterization has shown a maximum PDP of 46 % and a DCR of 0.4 Hz/μm^2 with and excess voltage of 3 V. All-in-all experimental results indicate that the employed 110 nm CIS technology is a convenient choice for further implementations.

- Photo-emission characterization has shown that PEB occurs in PP/NW structure with a non-uniform pattern, indicating the presence of non-regular doping. Regarding the PW/DNW structure, it presents a more uniform glow around the whole perimeter, also showing the existence

of PEB but at higher excess voltage. The glow at the border can be a consequence of the small guard ring width of 1 µm.

- Experimental measurements show that self-sustained avalanches may occur despite the AQRC being sized to stop them based on models and data available from the foundry. Such self-sustained avalanches would result into an experimental power consumption of ~ 40 W per SiPM, and precluded global characterization. At the conceptual level, the reason for such behavior might be that the AQRC relies on, first, detecting the avalanche and, then, switching off the quenching transistor. Because avalanche currents can grow to milliamperes, delays in performing the quenching action must be very short. To that purpose, the quenching transistor must have a minimum resistance that ensures a passive quenching at any time and PVT variation.

- During the design of the TDC based on VCROs, two delay stages were studied: XCP and XCI. The XCI stage provides higher oscillation frequency, while the XCP stage has a lower phase noise. In terms of TDC performance, XCI stages have a better FOM_{TDC} than XCP stages, with its smallest value at 0.3 pJ and 0.5 pJ per conversion, respectively.

- The results of the TDC characterization agree with post-layout simulations, showing a time resolution below 100 ps with a power consumption of 1.1 mW. The DNL is ±0.4 LSB, while the INL is ±6.1 LSB. In terms of TDC performance, there is still room for further improvements. On the one hand, the DNL and INL can be improved by using a PLL to compensate for PVT variations. On the other hand, experimental measurements confirm an expectable trade-off between area occupation and substrate noise coupling. The prototype chip targeted minimum area occupation, thus placing the ring oscillator very close to the rest of the logic. Closer assessment of the trade-off is advisable for smaller coupling between the ring oscillator and the ripple counter.

- Architectural studies explored the design space by following a bottom-up approach in the quest to determine the dependencies of parameters like sensitivity, SNR, or linearity, on the one hand, and the number of SPADs per micro-cell and the number of sub-pixels per pixel, on the other hand. Besides providing insight into the trade-offs, optimum settings were identified for prototype implementation.

BIBLIOGRAPHY

[1] W. Röntgen and L. Anna Bertha Ludwig; Wellcome Library, *Hand Mit Ringen, hand with ring*. [Online]. Available: https://wellcomecollection.org/works/wjc8ejn2

[2] S. Derenzo, M. Boswell, M. Weber, and K. Brennan. (2008) Scintillation properties. [Online]. Available: http://scintillator.lbl.gov/

[3] R. Mao, L. Zhang, and R.-Y. Zhu, "Optical and scintillation properties of inorganic scintillators in high energy physics," *Nuclear Science, IEEE Transactions on*, vol. 55, no. 4, pp. 2425–2431, Aug 2008.

[4] T. Kimble, M. Chou, and B. Chai, "Scintillation properties of LYSO crystals," in *Nuclear Science Symposium Conference Record, 2002 IEEE*, vol. 3, Nov 2002, pp. 1434–1437 vol.3.

[5] A. M. Thomas and A. K. Banerjee, *The History of Radiology*. Oxford University Press, 2013.

[6] M. Guarnieri, "The age of vacuum tubes: Early devices and the rise of radio communications [historical]," *IEEE Industrial Electronics Magazine*, vol. 6, no. 1, pp. 41–43, March 2012.

[7] S. Siddharth and A. Goyal, "The origin of echocardiography: a tribute to Inge Edler," *Texas Heart Institute journal*, vol. 34, no. 4, pp. 431–438, feb 2007.

[8] G. Watts, "John Wild," *BMJ*, vol. 339, 2009. [Online]. Available: https://www.bmj.com/content/339/bmj.b4428

[9] P. C. Lauterbur, "Image formation by induced local interactions: Examples employing nuclear magnetic resonance," *Nature*, vol. 242, no. 5394, pp. 190–191, 1973. [Online]. Available: https://doi.org/10.1038/242190a0

[10] M. M. Ter-Pogossian, M. E. Phelps, E. J. Hoffman, and N. A. Mullani, "A positron-emission transaxial tomograph for nuclear imaging PET," *Radiology*, vol. 114, no. 1, pp. 89–98, 1975, pMID: 1208874. [Online]. Available: https://doi.org/10.1148/114.1.89

[11] G. Gowrishankar, M. Namavari, E. B. Jouannot, A. Hoehne, R. Reeves, J. Hardy, and S. S. Gambhir, "Investigation of 6-[18-F]-fluoromaltose as a novel PET tracer for imaging bacterial infection," *PloS one*, vol. 9, no. 9, pp. e107 951–e107 951, Sep 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25243851

[12] K. Iniewski, *Medical Imaging: Principles, Detectors, and Electronics*. Wiley, 2009.

[13] W. W. Moses, "Fundamental limits of spatial resolution in PET," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 648, Supplement 1, pp. S236 – S240, 2011, NIMA 4th International Conference on Imaging techniques in Subatomic Physics, Astrophysics, Medicine, Biology and Industry. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900210026276

[14] M. Conti, "Focus on time-of-flight PET: the benefits of improved time resolution," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 38, no. 6, pp. 1147–1157, Jun 2011. [Online]. Available: https://doi.org/10.1007/s00259-010-1711-y

[15] D. A. López-Mora, A. Flotats, F. Fuentes-Ocampo, V. Camacho, A. Fernández, A. Ruiz, J. Duch, M. Sizova, A. Domènech, M. Estorch, and I. Carrió, "Comparison of image quality and lesion detection between digital and analog PET/CT," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 46, no. 6, pp. 1383–1390, 2019. [Online]. Available: https://doi.org/10.1007/s00259-019-4260-z

[16] (2018) Physical performance of first total-body explorer PET scanner and preclinical applications with mini-explorer systems. [Online]. Available: https://explorer.ucdavis.edu/

[17] C. W. E. van Eijk, "Inorganic scintillators in medical imaging," *Physics in Medicine and Biology*, vol. 47, no. 8, pp. R85–R106, apr 2002. [Online]. Available: https://doi.org/10.1088%2F0031-9155%2F47%2F8%2F201

[18] W. R. Leo, *Leo, W. R. (1994). Techniques for Nuclear and Particle Physics Experiments*. Springer, 1994.

[19] P. Lecoq, "Pushing the limits in time-of-flight PET imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 1, no. 6, pp. 473–485, Nov 2017.

[20] G. F. Knoll, *Radiation Detection and Measurement*, 3rd ed. John Wiley and Sons, 1999.

[21] F. Wilkinson, "Chapter 13 - Scintillators," in *Emission Tomography*, M. N. Wernick and J. N. Aarsvold, Eds. San Diego: Academic Press, 2004, pp. 229 – 254. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780127444826500168

[22] R. Mao, L. Zhang, and R.-Y. Zhu, "LSO/LYSO crystals for future HEP experiments," *Journal of Physics: Conference Series*, vol. 293, no. 1, p. 012004, 2011. [Online]. Available: http://stacks.iop.org/1742-6596/293/i=1/a=012004

[23] V. Ilisie, V. Giménez-Alventosa, L. Moliner, F. Sánchez, A. González, M. Rodríguez-Álvarez, and J. Benlloch, "Building blocks of a multi-layer pet with time sequence photon interaction discrimination and double compton camera," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 895, pp. 74 – 83, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900218304431

[24] R. W. Todd, J. M. Nightingale, and D. B. Everett, "A proposed gamma-camera," *Nature*, vol. 251, no. 5471, pp. 132–134, 1974. [Online]. Available: https://doi.org/10.1038/251132a0

[25] C. W. Lerche, J. M. Benlloch, F. Sanchez, N. Pavon, B. Escat, E. N. Gimenez, M. Fernandez, I. Torres, M. Gimenez, A. Sebastia, and J. Martinez, "Depth of Gamma-ray interaction within continuous crystals from the width of its scintillation light-distribution," *IEEE Transactions on Nuclear Science*, vol. 52, no. 3, pp. 560–572, June 2005.

[26] R. Marcinkowski, P. Mollet, R. V. Holen, and S. Vandenberghe, "Sub-millimetre DOI detector based on monolithic LYSO and digital SiPM for a dedicated small-animal PET system," *Physics in Medicine and Biology*, vol. 61, no. 5, pp. 2196–2212, feb 2016. [Online]. Available: https://doi.org/10.1088%2F0031-9155%2F61%2F5%2F2196

[27] A. J. González, A. Aguilar, P. Conde, L. Hernández, L. Moliner, L. F. Vidal, F. Sánchez, S. Sánchez, C. Correcher, C. Molinos, J. Barberá, K. Lankes, S. Junge, T. Bruckbauer, P. Bruyndonckx, and J. M. Benlloch, "A PET design based on SiPM and monolithic LYSO crystals: Performance evaluation," *IEEE Transactions on Nuclear Science*, vol. 63, no. 5, pp. 2471–2477, Oct 2016.

[28] A. Knapitsch and P. Lecoq, "Review on photonic crystal coatings for scintillators," *International Journal of Modern Physics A*, vol. 29, no. 30, p. 1430070, 2014. [Online]. Available: https://doi.org/10.1142/S0217751X14300701

[29] F. Gramuglia, M. Lee, E. Venialgo, C. Bruschini, and E. Charbon, "Towards 10ps SPTR and ultra-low DCR in SiPMs through the combination of microlenses and photonic crystals," in *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Oct 2017, pp. 1–3.

[30] B. Singh, M. S. J. Marshall, S. Waterman, C. Pina-Hernandez, A. Koshelev, K. Munechika, A. Knapitsch, M. Salomoni, R. Pots, P. Lecoq, and V. V. Nagarkar, "Enhanced scintillation light extraction using nanoimprinted photonic crystals," *IEEE Transactions on Nuclear Science*, vol. 65, no. 4, pp. 1059–1065, April 2018.

[31] H. Iams and B. Salzberg, "The secondary emission phototube," *Proceedings of the Institute of Radio Engineers*, vol. 23, no. 1, pp. 55–64, Jan 1935.

[32] A. Spinelli and A. L. Lacaita, "Physics and numerical simulation of single photon avalanche diodes," *IEEE Transactions on Electron Devices*, vol. 44, no. 11, pp. 1931–1943, Nov 1997.

[33] S. M. Sze and K. N. Kwok, *Physics of Semiconductor Devices*. John Wiley and Sons, 3rd Edition, 2006.

[34] A. C. Ulku, C. Bruschini, I. M. Antolovic, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512 x 512 SPAD image sensor with integrated gating for widefield FLIM," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, Jan 2019.

[35] C. Zhang, "CMOS SPAD sensor for 3D time-of-flight imaging, LiDAR, and ultra-high speed cameras," Ph.D. dissertation, PhD thesis, Delft University of Technology, 2019.

[36] A. Ximenes, "Modular time-of-flight image sensor for light detection and ranging," Ph.D. dissertation, PhD thesis, Delft University of Technology, 2019.

[37] (2018) Hamamatsu photonics: Multi-pixel photon counters. [Online]. Available: https://www.hamamatsu.com/us/en/product/optical-sensors/mppc/mppc_mppc-array/index.html

[38] (2018) SensL: Silicon photomultipliers and SPADs. [Online]. Available: https://www.onsemi.com/PowerSolutions/parametrics/17940/techdocs

[39] S. Dolinsky, G. Fu, and A. Ivan, "Timing resolution performance comparison for fast and standard outputs of SensL SiPM," in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, Oct 2013, pp. 1–6.

[40] T. Pro, A. Ferri, A. Gola, N. Serra, A. Tarolli, N. Zorzi, and C. Piemonte, "New developments of near-UV SiPMs at FBK," *IEEE Transactions on Nuclear Science*, vol. 60, no. 3, pp. 2247–2253, June 2013.

[41] J. P. Schmall, R. I. Wiener, S. Surti, A. Ferri, A. Gola, A. Tarolli, C. Piemonte, and J. S. Karp, "Timing and energy resolution of new near-UV SiPMs coupled to $LaBr_3$:Ce for TOF-PET," *IEEE Transactions on Nuclear Science*, vol. 61, no. 5, pp. 2426–2432, Oct 2014.

[42] (2018) Philips: Digital photon counting: Silicon photomultiplier. [Online]. Available: https://www.usa.philips.com/healthcare/sites/digital-pet-learning-center

[43] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier principle of operation and intrinsic detector performance," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, Oct 2009, pp. 1959–1965.

[44] L. Braga, L. Gasparini, L. Grant, R. Henderson, N. Massari, M. Perenzoni, D. Stoppa, and R. Walker, "A fully digital 8x16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *Solid-State Circuits, IEEE Journal of*, vol. 49, no. 1, pp. 301–314, Jan 2014.

[45] S. Mandai, "Multichannel digital silicon photomultiplier for time-of-flight PET," Ph.D. dissertation, PhD thesis, Delft University of Technology, 2014.

[46] C. Xu, "Study of the silicon photomultipliers and their applications in positron emission tomography," Ph.D. dissertation, PhD thesis, University of Hamburg, 2014.

[47] M. W. Fishburn, "Fundamentals of CMOS single-photon avalanche diodes," Ph.D. dissertation, PhD thesis, Technische Universiteit Delft, 2012.

[48] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable single-photon avalanche diode structures in nanometer CMOS technology," *IEEE Transactions on Electron Devices*, vol. 58, no. 7, pp. 2028–2035, July 2011.

[49] E. A. G. Webster, J. A. Richardson, L. A. Grant, and R. Henderson, "An infra-red sensitive, low noise, single-photon avalanche diode in 90nm CMOS," in *2011 International Image Sensor Workshop, Japan*, Jun 2011, pp. 102–105, paper.

[50] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions," *Applied Physics Letters*, vol. 57, no. 5, pp. 489–491, 1990.

[51] R. Newman, "Visible light from a silicon $p-n$ junction," *Phys. Rev.*, vol. 100, pp. 700–703, Oct 1955. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.100.700

[52] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," *Applied Physics Letters*, vol. 62, no. 6, pp. 606–608, 1993. [Online]. Available: https://doi.org/10.1063/1.108870

[53] C. L. Niclass, "Single-photon image sensors in CMOS: Picosecond resolution for three-dimensional imaging," Ph.D. dissertation, PhD thesis, Ecole Polytechnique Federale de Lausanne, 2008.

[54] H. Xu, L. Pancheri, L. H. C. Braga, G.-F. D. Betta, and D. Stoppa, "Crosstalk characterization of dense single-photon avalanche diode arrays in CMOS 150nm technology," *Optical Engineering*, vol. 55, no. 6, pp. 1 – 5 – 5, 2016. [Online]. Available: https://doi.org/10.1117/1.OE.55.6.067102

[55] H. Xu, L. Pancheri, G.-F. D. Betta, and D. Stoppa, "Design and characterization of a p+/n-well SPAD array in 150nm CMOS process," *Opt. Express*, vol. 25, no. 11, pp. 12 765–12 778, May 2017. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-11-12765

[56] M. Moreno-García, H. Xu, L. Gasparini, and M. Perenzoni, "Low-noise single photon avalanche diodes in a 110nm CIS technology," in *2018 48th European Solid-State Device Research Conference (ESSDERC)*, Sep. 2018, pp. 94–97.

[57] N. V. Loukianova, H. O. Folkerts, J. P. V. Maas, D. W. E. Verbugt, A. J. Mierop, W. Hoekstra, E. Roks, and A. J. P. Theuwissen, "Leakage current modeling of test structures for characterization of dark current in CMOS image sensors," *IEEE Transactions on Electron Devices*, vol. 50, no. 1, pp. 77–83, Jan 2003.

[58] H. Finkelstein, M. J. Hsu, and S. C. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology," *IEEE Electron Device Letters*, vol. 27, no. 11, pp. 887–889, Nov 2006.

[59] M. Hsu, S. Esener, and H. Finkelstein, "A CMOS STI-bound single-photon avalanche diode with 27ps timing resolution and a reduced diffusion tail," *Electron Device Letters, IEEE*, vol. 30, no. 6, pp. 641–643, June 2009.

[60] M. M. Vignetti, F. Calmon, R. Cellier, P. Pittet, L. Quiquerez, and A. Savoy-Navarro, "Design guidelines for the integration of geiger-mode avalanche diodes in standard CMOS technologies," *Microelectronics Journal*, vol. 46, no. 10, pp. 900–910, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026269215001706

[61] A. L. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," *IEEE Transactions on Electron Devices*, vol. 40, no. 3, pp. 577–582, March 1993.

[62] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajic, G. O. S. Williams, J. M. Girkin, and R. K. Henderson, "A CMOS SPAD line sensor with per-pixel histogramming TDC for time-resolved multispectral imaging," *IEEE Journal of Solid-State Circuits*, pp. 1–15, 2019.

[63] L. Pancheri and D. Stoppa, "Low-noise single photon avalanche diodes in 0.15μm CMOS technology," in *2011 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, Sep. 2011, pp. 179–182.

[64] P. R. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, June 2005.

[65] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, Apr 1996. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-35-12-1956

[66] G.-F. D. Betta, L. Pancheri, D. Stoppa, R. Henderson, and J. Richardson, "Avalanche photodiodes in submicron CMOS technologies for high-sensitivity imaging," in *Advances in Photodiodes*, G. F. D. Betta, Ed. Rijeka: IntechOpen, 2011, ch. 11. [Online]. Available: https://doi.org/10.5772/15178

[67] H. Finkelstein, M. J. Hsu, S. Zlatanovic, and S. Esener, "Performance trade-offs in single-photon avalanche diode miniaturization," *Review of Scientific Instruments*, vol. 78, no. 10, 2007. [Online]. Available: http://scitation.aip.org/content/aip/journal/rsi/78/10/10.1063/1.2796146

[68] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Transactions on Nuclear Science*, vol. 57, no. 6, pp. 3815–3826, Dec 2010.

[69] D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi, and F. Zappa, "Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects," *IEEE Photonics Technology Letters*, vol. 25, no. 8, pp. 776–779, April 2013.

[70] E. Panina, "Design and characterization of SPAD based CMOS analog pixels for photon-counting applications," Ph.D. dissertation, PhD thesis, Department of Information Engineering and Computer Science, University of Trento, 2014.

[71] G. Boso, M. Buttafava, F. Villa, and A. Tosi, "Low-cost and compact single-photon counter based on a CMOS SPAD smart pixel," *IEEE Photonics Technology Letters*, vol. 27, no. 23, pp. 2504–2507, Dec 2015.

[72] F. Zappa, A. Tosi, A. D. Mora, and S. Tisa, "Spice modeling of single photon avalanche diodes," *Sensors and Actuators A: Physical*, vol. 153, no. 2, pp. 197–204, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924424709002581

[73] E. Kamrani, F. Lesage, and M. Sawan, "Efficient premature edge breakdown prevention technique for SiAPD fabrication using standard CMOS process," *IOP International Journal of Semiconductor Science and Technology (JSST)*, vol. 28, no. 4, p. 045008, 2013. [Online]. Available: http://iopscience.iop.org/0268-1242/28/4/045008

[74] S. Tisa, F. Guerrieri, and F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes," *Opt. Express*, vol. 16, no. 3, pp. 2232–2244, Feb 2008. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-16-3-2232

[75] M. Gronholm, J. Poikonen, and M. Laiho, "A ring-oscillator-based active quenching and active recharge circuit for single photon avalanche diodes," in *2009 European Conference on Circuit Theory and Design*, Aug 2009, pp. 5–8.

[76] A. Rochas, "Single photon avalanche diodes in CMOS technology," Ph.D. dissertation, PhD thesis, Ecole Polytechnique Federale de Lausanne, 2003.

[77] I. Vornicu, R. Carmona-Galán, B. Pérez-Verdú, and A. Rodríguez-Vázquez, "Compact CMOS active quenching/recharge circuit for SPAD arrays," *International Journal of Circuit Theory and Applications*, vol. 44, no. 4, pp. 917–928, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cta.2113

[78] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-PDE, backside-illuminated SPAD in 65/40nm 3D IC CMOS pixel with cascoded passive quenching and active recharge," *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547–1550, Nov 2017.

[79] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon, "Nonuniformity analysis of a 65kpixel CMOS SPAD imager," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 57–64, Jan 2016.

[80] D. Portaluppi, E. Conca, and F. Villa, "32 x 32 CMOS SPAD imager for gated imaging, photon timing, and photon coincidence," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–6, March 2018.

[81] L. H. C. Braga, L. Pancheri, L. Gasparini, M. Perenzoni, R. Walker, R. K. Henderson, and D. Stoppa, "A CMOS mini-SiPM detector with in-pixel data compression for PET applications," in *2011 IEEE Nuclear Science Symposium Conference Record*, Oct 2011, pp. 548–552.

[82] S. Henzler, *Time-to-digital Converters*. Springer, 2010.

[83] M. Z. Straayer and M. H. Perrott, "A multi-path gated ring oscillator TDC with first-order noise shaping," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1089–1098, April 2009.

[84] A. A. Hafez and C. K. Yang, "Design and optimization of multipath ring oscillators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 10, pp. 2332–2345, Oct 2011.

[85] M. Lee and A. A. Abidi, "A 9b, 1.25ps resolution coarse-fine time-to-digital converter in 90nm CMOS that amplifies a time residue," in *2007 IEEE Symposium on VLSI Circuits*, June 2007, pp. 168–169.

[86] J. S. Kim, Y. H. Seo, Y. Suh, H. J. Park, and J. Y. Sim, "A 300MS/s, 1.76ps resolution, 10b asynchronous pipelined time-to-digital converter with on-chip digital background calibration in 0.13μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 516–526, Feb 2013.

[87] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, "A local passive time interpolation concept for variation-tolerant high-resolution time-to-digital conversion," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666–1676, July 2008.

[88] I. Vornicu, R. Carmona-Galán, and Rodríguez-Vázquez, "In-pixel voltage-controlled ring-oscillator for phase interpolation in ToF image sensors," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2016, pp. 1906–1909.

[89] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, "A 32x32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," in *2009 IEEE Custom Integrated Circuits Conference*. IEEE Solid-State Circuits Society, Sept 2009, pp. 77–80.

[90] Z. Cheng, X. Zheng, M. J. Deen, and H. Peng, "Recent developments and design challenges of high-performance ring oscillator CMOS time-to-digital converters," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 235–251, Jan 2016.

[91] J. A. McNeill and D. Ricketts, *The Designer's Guide to Jitter in Ring Oscillators*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[92] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, "100 000 Frames/s 64x32 single-photon detector array for 2D imaging and 3D ranging," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 354–363, Nov 2014.

[93] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64×64-pixels digital silicon photomultiplier direct TOF sensor with 100-MPhotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6km for spacecraft navigation and landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, Jan 2017.

[94] A. Akindinov and et al., "Performance of the ALICE time-of-flight detector at the LHC," *The European Physical Journal Plus*, vol. 128, no. 4, p. 44, 2013. [Online]. Available: https://doi.org/10.1140/epjp/i2013-13044-x

[95] D. Tyndall, B. R. Rae, D. D. Li, J. Arlt, A. Johnston, J. A. Richardson, and R. K. Henderson, "A high-throughput time-resolved mini-silicon photomultiplier with embedded fluorescence lifetime estimation in 0.13μm CMOS," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 6, pp. 562–570, Dec 2012.

[96] J. Jalil, M. B. I. Reaz, and M. A. M. Ali, "CMOS differential ring oscillators: Review of the performance of CMOS ROs in communication systems," *IEEE Microwave Magazine*, vol. 14, no. 5, pp. 97–109, July 2013.

[97] B. Razavi, *Design of Monolithic Phase Locked Loops and Clock Recovery Circuits: A Tutorial*. IEEE, 1996. [Online]. Available: https://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5263316

[98] W. Bae, H. Ju, K. Park, S. Y. Cho, and D. K. Jeong, "A 7.6mW, 414fs RMS-jitter 10GHz phase-locked loop for a 40Gb/s serial link transmitter based on a two-stage ring oscillator in 65nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 10, pp. 2357–2367, Oct 2016.

[99] I. Vornicu, R. Carmona-Galán, and Rodríguez-Vázquez, "Arrayable voltage-controlled ring-oscillator for direct time-of-flight image sensors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. PP, no. 99, pp. 1–14, 2017.

[100] H. Chen and R. Geiger, "Transfer characterization of CMOS ring voltage controlled oscillators," in *Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems. MWSCAS 2001 (Cat. No.01CH37257)*, vol. 1. IEEE Circuits and Systems Society, 2001, pp. 66–70 vol.1.

[101] I. Nissinen, A. Mantyniemi, and J. Kostamovaara, "A CMOS time-to-digital converter based on a ring oscillator for a laser radar," in *ESSCIRC 2004 - 29th European Solid-State Circuits Conference (IEEE Cat. No.03EX705)*. IEEE Solid-State Circuits Society, Sept 2003, pp. 469–472.

[102] M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R. Henderson, F. Borghetti, D. Stoppa, and E. Charbon, "A parallel 32x32 time-to-digital converter array fabricated in a 130nm imaging CMOS technology," in *2009 Proceedings of ESSCIRC*. ESSCIRC, Sept 2009, pp. 196–199.

[103] Z. Cheng, M. J. Deen, and H. Peng, "A low-power gateable vernier ring oscillator time-to-digital converter for biomedical imaging applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp. 445–454, April 2016.

[104] C. Veerappan, J. Richardson, R. Walker, D. U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160x128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *2011 IEEE International Solid-State Circuits Conference*. IEEE Solid-State Circuits Society, Feb 2011, pp. 312–314.

[105] M. Perenzoni, L. Gasparini, and D. Stoppa, "Design and characterization of a 43.2ps and PVT-resilient TDC for single-photon imaging arrays," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 4, pp. 411–415, April 2018.

[106] B. Razavi, "A study of phase noise in CMOS oscillators," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 3, pp. 331–343, Mar 1996.

[107] A. Hajimiri and T. H. Lee, "A general theory of phase noise in electrical oscillators," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 2, pp. 179–194, Feb 1998.

[108] A. Hajimiri, S. Limotyrakis, and T. H. Lee, "Jitter and phase noise in ring oscillators," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 6, pp. 790–804, June 1999.

[109] A. A. Abidi, "Phase noise and jitter in CMOS ring oscillators," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 8, pp. 1803–1816, Aug 2006.

[110] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor*. Oxford University Press, 3rd Edition, 2011.

[111] J. Chang, A. A. Abidi, and C. R. Viswanathan, "Flicker noise in CMOS transistors from subthreshold to strong inversion at various temperatures," *IEEE Transactions on Electron Devices*, vol. 41, no. 11, pp. 1965–1971, Nov 1994.

[112] D. Xie, M. Cheng, and L. Forbes, "SPICE models for flicker noise in n-MOSFETs from subthreshold to strong inversion," *IEEE Transactions on*

Computer-Aided Design of Integrated Circuits and Systems, vol. 19, no. 11, pp. 1293–1303, Nov 2000.

[113] L. K. J. Vandamme and F. N. Hooge, "What do we certainly know about $1/f$ noise in mosts?" *IEEE Transactions on Electron Devices*, vol. 55, no. 11, pp. 3070–3085, Nov 2008.

[114] K. Kim, Y. H. Kim, W. Yu, and S. Cho, "A 7bit, 3.75ps resolution two-step time-to-digital converter in 65nm CMOS using pulse-train time amplifier," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 1009–1017, April 2013.

[115] P. Napolitano, A. Moschitta, and P. Carbone, "A survey on time interval measurement techniques and testing methods," in *2010 IEEE Instrumentation Measurement Technology Conference Proceedings*. IEEE Instrumentation Measurement Technology Conference Proceedings, May 2010, pp. 181–186.

[116] I. Vornicu, R. Carmona-Galán, and Rodríguez-Vázquez, "On the calibration of a SPAD-based 3D imager with in-pixel TDC using a time-gated technique," in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE Circuits and Systems Society, May 2015, pp. 1102–1105.

[117] N. Zorzi, M. Melchiorri, A. Piazza, C. Piemonte, and A. Tarolli, "Development of large-area silicon photomultiplier detectors for PET applications at FBK," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 636, no. 1, Supplement, pp. S208 – S213, 2011, 7th International Hiroshima Symposium on the Development and Application of Semiconductor Tracking Detectors. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168900210009654

[118] M. W. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillators," *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, pp. 2549–2557, Oct 2010.

[119] M. Tétrault, A. C. Therrien, W. Lemaire, R. Fontaine, and J. Pratte, "TDC array tradeoffs in current and upcoming digital SiPM detectors for Time-of-Flight PET," *IEEE Transactions on Nuclear Science*, vol. 64, no. 3, pp. 925–932, March 2017.

[120] F. Acerbi, A. Gola, V. Regazzoni, G. Paternoster, G. Borghi, N. Zorzi, and C. Piemonte, "High efficiency, ultra-high-density silicon photomultipliers," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–8, March 2018.

[121] N. A. W. Dutton, S. Gnecchi, L. Parmesan, A. J. Holmes, B. Rae, L. A. Grant, and R. K. Henderson, "11.5 a time-correlated single-photon-counting sensor with 14GS/S histogramming time-to-digital converter," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, Feb 2015, pp. 1–3.

[122] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. Pellegrini, S. J. McLeod, L. A. Grant, and R. K. Henderson, "Digital silicon photomultipliers with OR/XOR pulse combining techniques," *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 1105–1110, March 2016.

[123] M. A. Karami, "Deep-submicron CMOS single photon detectors and quantum effects," Ph.D. dissertation, PhD thesis, Delft University of Technology, 2011.

[124] C. Veerappan and E. Charbon, "A low dark count p-i-n diode based SPAD in CMOS technology," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 65–71, Jan 2016.

[125] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, "A low-noise single-photon detector implemented in a 130nm CMOS imaging process," *Solid-State Electronics*, vol. 53, no. 7, pp. 803–808, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038110109000896

[126] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500μm diameter and 55% detection efficiency at 420 nm," *Journal of Modern Optics, Volume 61, Issue 2, 2014, pp.102-115*, vol. 61, pp. 102–115, Jan 2014. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2014JMOp...61..102V

[127] E. A. G. Webster and R. K. Henderson, "A TCAD and spectroscopy study of dark count mechanisms in single-photon avalanche diodes," *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4014–4019, Dec 2013.

[128] I. Vornicu, F. Bandi, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "Low-noise and high-efficiency near-IR SPADs in 110nm CIS technology," in *2019 49th European Solid-State Device Research Conference (ESSDERC)*, Sep. 2019, pp. 250–253.

[129] N. D'Ascenzo and Q. Xie, "Possible layout solutions for the improvement of the dark rate of Geiger mode avalanche structures in the GLOBALFOUNDRIES BCDLITE 0.18 um CMOS technology," *Journal of Instrumentation*, vol. 13, no. 04, pp. T04 007–T04 007, apr 2018. [Online]. Available: https://doi.org/10.1088/1748-0221/13/04/T04007

SCIENTIFIC PUBLICATIONS

In the following, the list of contributions to the scientific bibliography made in this dissertation is provided.

Workshops

[1] F. N. Bandi, I. Vornicu, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "VCRO-Based TDcs in Submicron CIS Technology," in *6th Workshop on Architecture of Smart Cameras (WASC)*, Jun. 5-6, 2017, Córdoba, Spain.

[2] F. N. Bandi, I. Vornicu, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "Digital Silicon Photomultiplier Design Flow," in *7th Workshop on Architecture of Smart Cameras (WASC)*, Jun. 27-28, 2018, Coimbra, Portugal.

Seminars

[1] F. N. Bandi, "Single Photon Detection in CMOS technologies," in *1st Physics Doctoral Meeting of the Colegio Doctoral de Tordesillas*, Nov. 11-13, 2017, Granada, Spain.

Conferences

[1] F. N. Bandi, I. Vornicu, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "Design of a compact and low-power TDC for an array of SiPM's in 110nm CIS technology," in *2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, Jun. 12-15, 2017, Giardini Naxos, Italy, pp. 257-260.

[2] I. Vornicu, F. N. Bandi, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "Low-Noise and High-Efficiency Near-IR SPADs in 110nm CIS Technology," in *2019 49th European Solid State Device Research Conference (ESSDERC)*, Sept. 23-26, 2019, Krakow, Poland, pp. 250-253.

Journals

[1] I. Vornicu, F. N. Bandi, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "A CMOS Digital SiPM With Focal-Plane Light-Spot Statistics for DOI Computation," in *IEEE Sensors Journal*, vol. 17, no. 3, pp. 632-643, 1 Feb.1, 2017.

Under Review

[1] R. Ballabriga, J. Alozy, F. N. Bandi, M. Campbell, N. Egidos, J. M. Fernandez-Tenllado, E.H.M. Heijne, Fellow, IEEE, I. Kremastiotis, X. Llopart, B. J. Madsen, D. Pennicard, V. Sriskaran and L. Tlustos, "Photon Counting Detectors for X-ray Imaging with Emphasis on CT," in *IEEE Transactions on Radiation and Plasma Medical Sciences*.

[1] I. Vornicu, F. N. Bandi, R. Carmona-Galán, and Ángel Rodríguez-Vázquez, "Compact Macro-Cell with OR Pulse Combining for Low Power Digital-SiPM," in *IEEE Sensors Journal*.